Lecture 1: Introduction to Linear Regression

Contents

Contributors: Gabriela Cohen Freue, Katie Burak

High-level goals of this course:

- Build a strong conceptual and practical understanding of linear regression
- Become competent using `R` to fit and use a linear models to analyze real data
- Distinguish between generative models (GM) and predictive models (PM):
  - GM: to understand the data-generating mechanism
  - PM: to make out-of-sample predictions using the estimated model
- Learn how to interpret results and measure their uncertainty

Limitations:

- Linear Regression is a rich field and we won't be able to exhaustively cover it in 8 lectures
- Mathematical derivations are mostly omitted in lectures but some additional details are given in supplemental reading material
- Inference will be restricted to particular cases

Lecture 1: Learning Objectives

By the end of this lecture you will be able to:

- Formulate a Linear Regression (LR)
- Compare different definitions of LR
- Identify different goals of a LR
- Identify different elements and assumptions of a LR
- Define common terminology

References:

- Part of the content of this lecture follows closely Chapter 17 of Introduction to Data Science, Rafael Irizarry (book available at leanpub)

How have linear regression models been used to date?

As data scientists, we are often interested in understanding the relationship between variables in our data using *models*.

For example:

- which variables are associated with a response of interest?
- can we model the relationship between the response and the input variables? Is a linear model adequate?
- which variables are positively/negatively associated with the response?
- does the relationship between the response and an input variable depend on the values of the other variables?

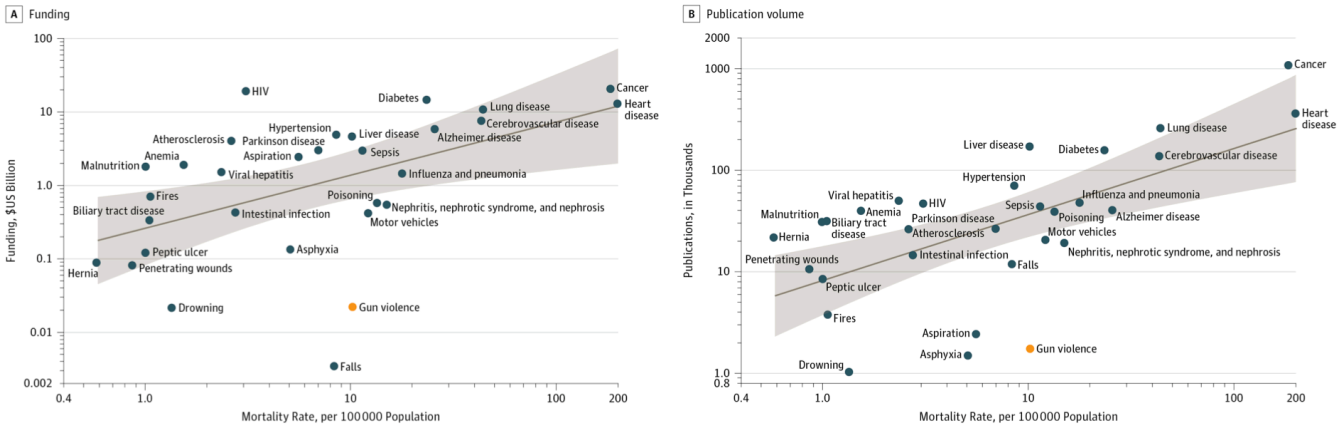
1.1 Examples

Sports: an example of prediction

Billy Bean, manager of the Oakland Athletics, used statistics to identify low cost players who can help the team win (example from Introduction to Data Science, Rafael Irizarry)

Public Health: an example of estimation

[Funding and Publication of Research on Gun Violence]



Reference: *JAMA* 2017; 317(1):84-85

Fetured twice in New York Times: [Gun Research Is Suddenly Hot], [There's an Awful Lot We Still Don't Know About Guns]

Climate Change: an example of inference

[Here's The Best Place To Move If You're Worried About Climate Change]

Reference: featured article in FiveThirtyEight

Climate Amenities, Climate Change, and American Quality of Life

Economists have used different **linear regression models** to explain people's choices in relation to climate variables. Reference: *JAERE* 2016; 3(1): 205-246

Epidemiology: an example of estimation and inference

Protein Consumption and Bone Mineral Density in the Elderly : The Rancho Bernardo Study

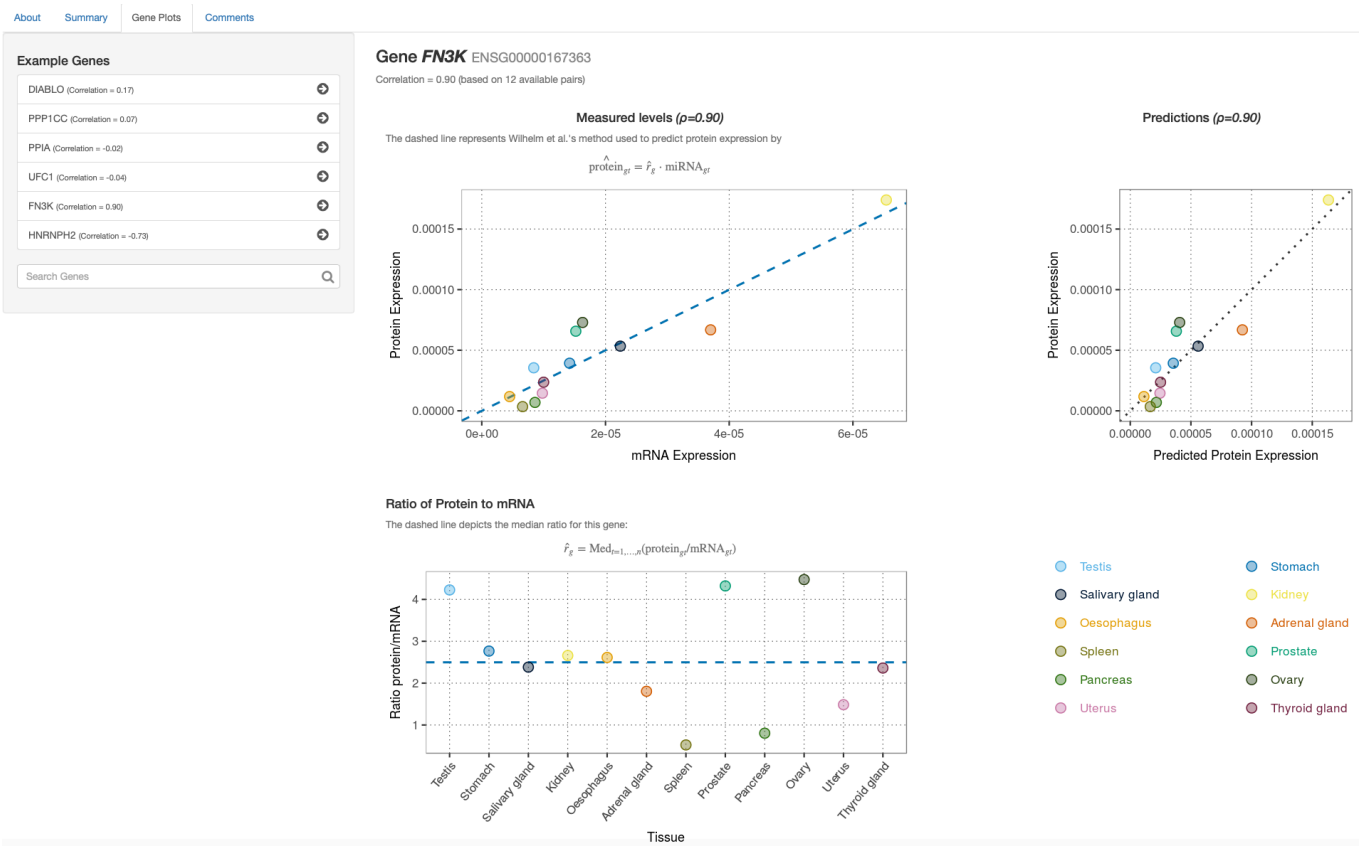
TABLE 2.
Results from linear regression analysis* of bone mineral density (g/cm^2) measured in 1992–1996 versus dietary protein variables measured in 1988–1992 for participants in the Rancho Bernardo Study

Variable	Total hip (*n* = 562 women, *n* = 382 men)†			Femoral neck (*n* = 562 women, *n* = 382 men)			Total spine (*n* = 568 women, *n* = 387 men)			Total body (*n* = 536 women, *n* = 369 men)		
	Estimated β coefficient	95% CI‡	*p* value	Estimated β coefficient	95% CI	*p* value	Estimated β coefficient	95% CI	*p* value	Estimated β coefficient	95% CI	*p* value
Women												
Protein§ (15 g)	0.0143	−0.0026, 0.0261	0.02	0.0092	−0.0007, 0.0192	0.07	0.0150	−0.0020, 0.0320	0.08	0.0158	0.0058, 0.0257	0.002
Animal protein (15 g)	0.0182	−0.0070, 0.0293	0.001	0.0125	−0.0032, 0.0219	0.01	0.0215	−0.0054, 0.0375	0.01	0.0161	−0.0067, 0.0254	0.001
Vegetable protein (5 g)	−0.0097	−0.0184, −0.0009	0.03	−0.0078	−0.0151, −0.0005	0.04	−0.0147	−0.0273, −0.0021	0.02	−0.0039	−0.0114, 0.0037	0.31
Men												
Protein (15 g)	0.0057	−0.0102, 0.0217	0.48	0.0032	−0.0110, 0.0174	0.66	0.0057	−0.0172, 0.0286	0.62	−0.0005	−0.0127, 0.0117	0.94
Animal protein (15 g)	0.0102	−0.0054, 0.0259	0.20	0.0059	−0.0081, 0.0199	0.41	0.0120	−0.0104, 0.0345	0.29	0.0016	−0.0104, 0.0136	0.80
Vegetable protein (5 g)	−0.0108	−0.0244, 0.0029	0.12	−0.0065	−0.0187, 0.0057	0.30	−0.0152	−0.0349, 0.0046	0.13	−0.0048	−0.0154, 0.0058	0.38

> "**Multiple linear regression analyses** adjusted for standard osteoporosis covariates showed a positive association between animal protein consumption, assessed by food frequency questionnaires in 1988–1992, and BMD, measured 4 years later." Reference: *Am J Epidemiol* 2002; 155(7):636–644

Medicine and Molecular Biology: an example of prediction

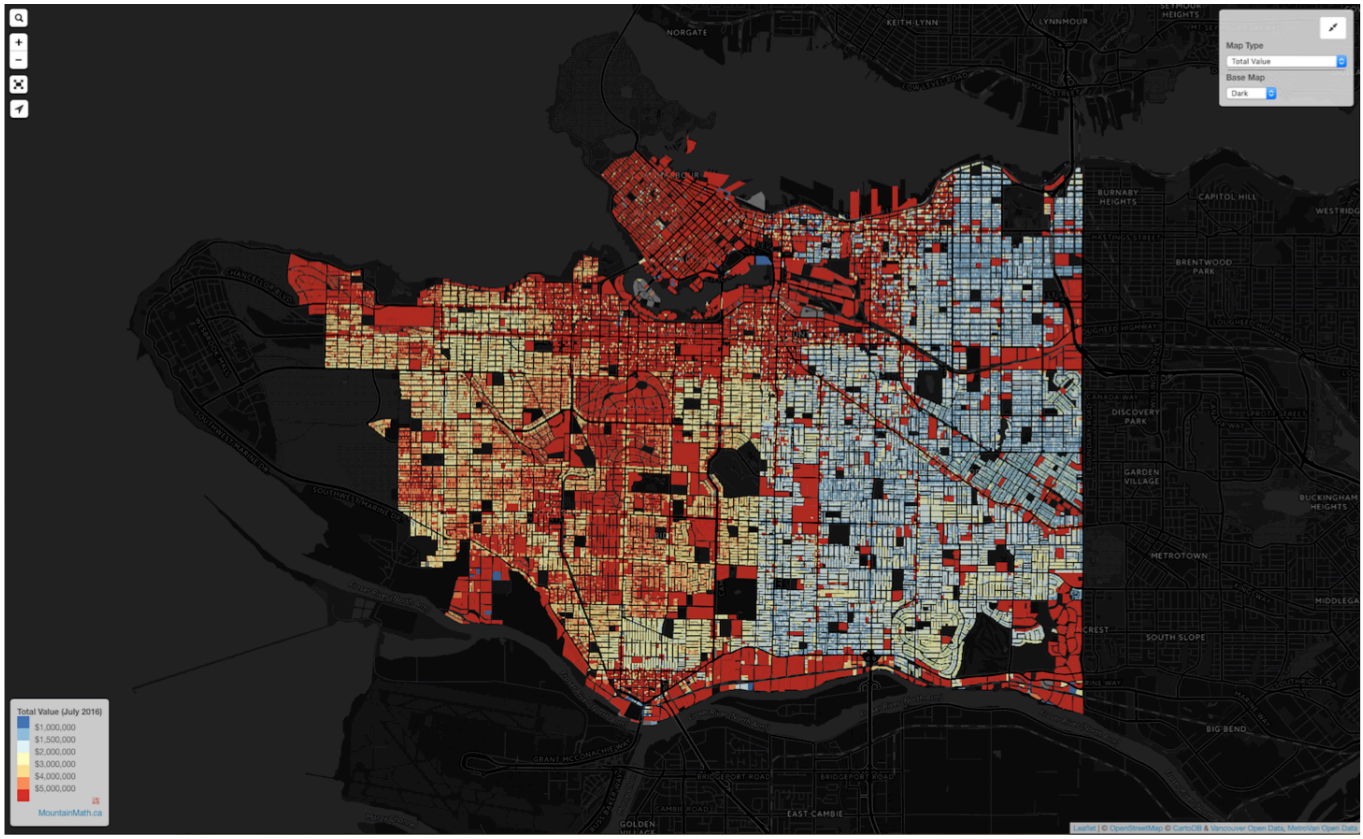
Can We Predict Protein from mRNA levels?



Reference: *Nature* 2017, 547:E19–E20

Real state: our working case study

Property Assessment Tax Data (Map produced by MountainMath)



We want to identify factors that determine the tax value of a property





1.2 Case Study: Property Tax Assessment

Class discussion:

Which variables are associated with property assessed value?

- neighbourhood (location)
- number of rooms
- square feet (size)
- age of house

Property Tax Assessment Case and Linear Regression

- Different type of variables may be associated with a property assessed value
- In DSCI_552 (Statistical Inference) you have learned tools to study the relation between a continuous and a categorical variable. For example, *do modern houses have a higher value than old houses?* You can:
  - use a *t*-test or a permutation test to test if the average value of modern houses is the same as that of old houses
- However, there are some questions that you won't be able to answer with the tools learned in DSCI_552:
- For example,
  - *is the assessed value associated with the size of the house?* (unless you collapse the information into 2 groups (large vs small houses)
  - if I build a new house, what would be its expected assessed value?

Conclusions

- Linear Models provide a unifying framework to estimate and test the true relation between different type of variables and the assessed value
- Linear Models can also be used to predict the assessed value of a house although it may not be the best predictive model

1.3 Estimation *vs.* inference *vs.* prediction

- **Estimation**: how to estimate the true (but unknown) relation between the response and the input variables
- **Inference**: how to use the model to infer information about the unknown relation between variables
- **Prediction**: how to use the model to predict the value of the response for new observations

Note: *These goals are related!*

Class Discussion:

Inference (I) or Prediction (P)?

Indicate which of the following are related to estimation and which are related to prediction:

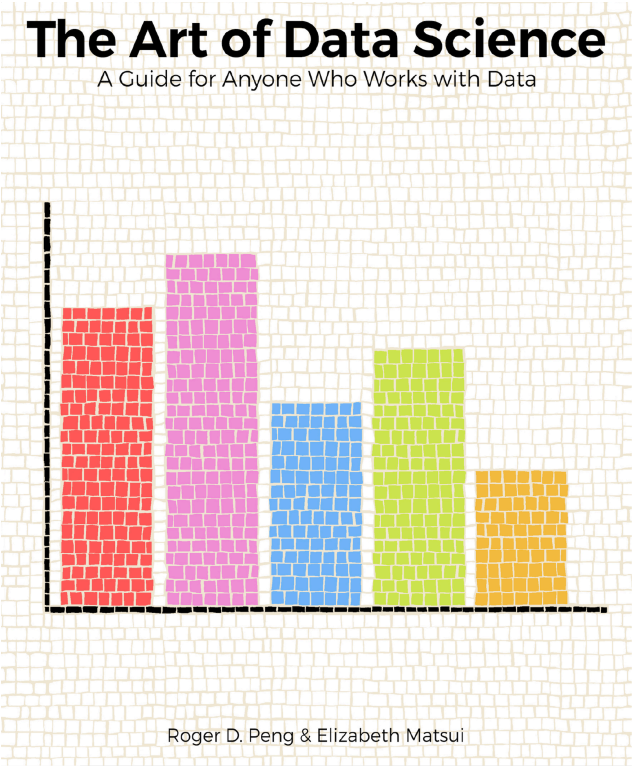
- (a) A real estate agent wants to identify factors that are related to the assessed values of homes (e.g., size of houses, age, amenities, etc)
- (b) A real estate agent is interested in determining if a house is under- or over-valued given its characteristics.
- (c) Biologists want to use mRNA data to predict protein values
- (d) Biologists want to verify empirically the central dogma of biology that relates mRNA to protein values
- (e) A research group wants to determine if wearing sunscreen is associated with a decreased probability of developing skin cancer in Canada.
- (f) A marketing team thinks that tweets with GIFs get on average more impressions than tweets that do not. They will conduct an A/B testing experiment to test their hypothesis.
- (g) A marketing team wants to forecast the number of mentions of a new tweet.

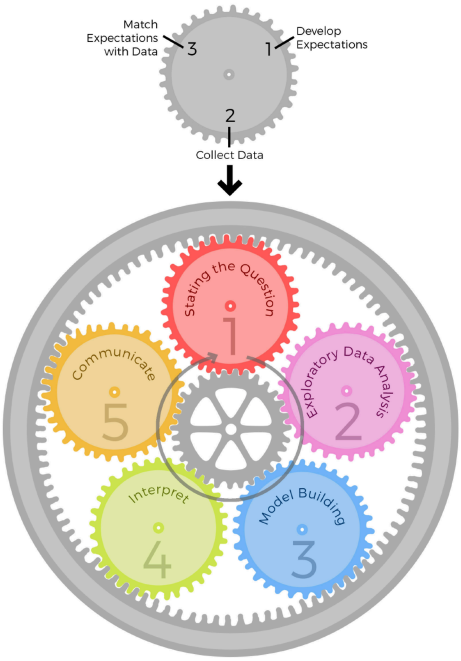
An important step in any data analysis is to explore and know important characteristics of the data. This includes, but it is not limited to:

- knowing the size of the data
- examining distributions of all variables using graphical and numerical summaries
- identifying missing values and potential outliers
- beginning to discover relationships between variables

This last step, usually referred as **Exploratory Data Analysis (EDA)**, is generally the first step in the analysis. However, a typical data science workflow is never linear and you may need to continue exploring the data at many points of the analysis path.

> **Heads-up**: Professors Peng and Matsui in their book "The Art of Data Science" described this process with **epicycles**





Epicycles of Analysis

2.1 EDA Checklist

1. Formulate your question
2. Read in your data
3. Look at the top and the bottom of your data
4. Check the packaging
5. Check your "n"s

> From The Art of Data Science, by Peng and Matsui (in a slightly different order)

The question:

We can use data to examine the relationship between the property assessed value and different characteristics of a property.

We will use a dataset that contains the property tax assessed values of properties in the Strathcona County. A valuation date of July 1, 2014 and a property condition date as of December 31, 2014 are provided.

Read in the data

Dataset: [2015 Property Tax Assessment from Strathcona County]

```
library(tidyverse)
library(broom)
library(repr)
library(ggplot2)
library(gridExtra)

dat <- read.csv("data/Assessment_2015.csv")

head(dat, n = 20)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ─
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2
```

```
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ─
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all con
```

```
Attaching package: ‘gridExtra’
```

```
The following object is masked from ‘package:dplyr’:

    combine
```

	the_geom	TAX_YEAR	ROLL_NUM	ADDRESS	YEAR_BUILT	AS
	<chr>	<int>	<dbl>	<chr>	<int>	
1	POINT (-112.9144294242738 53.447001849778545)	2015	1032301002		1940	
2	POINT (-112.87633420362904 53.41074767160115)	2015	1022705006		1950	Re
3	POINT (-112.91925958091502 53.40590216915645)	2015	1017328004	229 51268 RNG RD 204	1983	Re
4	POINT (-112.86295897279041 53.447144538217024)	2015	1034103000		1981	
5	POINT (-112.91933857223259 53.405754605523676)	2015	1017329002	231 51268 RNG RD 204	1967	
6	POINT (-112.9194226653068 53.40559810142708)	2015	1017330000	233 51268 RNG RD 204	1968	
7	POINT (-112.87706656864435 53.4105316732081)	2015	1015300005		1974	
8	POINT (-112.91921628521787 53.40720932431589)	2015	1017319003	205 51268 RNG RD 204	1962	Re
9	POINT (-112.92009423892009 53.40723973395918)	2015	1017322007	211 51268 RNG RD 204	1967	
10	POINT (-112.91926686738579 53.406126899561784)	2015	1017327006	227 51268 RNG RD 204	1983	Re
11	POINT (-112.91946041239973 53.405432191305344)	2015	1017331008	235 51268	1965	Re

	the_geom	TAX_YEAR	ROLL_NUM	ADDRESS	YEAR_BUILT	AS
	<chr>	<int>	<dbl>	<chr>	<int>	
				RNG RD 204		
12	POINT (-112.92120223308287 53.40413140696102)	2015	1017339001	253 51268 RNG RD 204	1969	
13	POINT (-112.91453939437695 53.408501408843904)	2015	1017304005	171 51268 RNG RD 204	1993	Re
14	POINT (-112.92142562194445 53.40425474383107)	2015	1017340009	255 51268 RNG RD 204	1964	
15	POINT (-112.92164872423655 53.4043848176214)	2015	1017341007	257 51268 RNG RD 204	1972	
16	POINT (-112.92005875795091 53.40508263450174)	2015	1017334002	241 51268 RNG RD 204	1962	
17	POINT (-112.92025810764135 53.404967611774744)	2015	1017335009	243 51268 RNG RD 204	1968	
18	POINT (-112.92045734252574 53.40485708096862)	2015	1017336007	245 51268 RNG RD 204	2005	Re
19	POINT (-112.92065657262883 53.40474205649951)	2015	1017337005	247 51268 RNG RD 204	1965	
20	POINT (-112.92085586227155 53.40462927888298)	2015	1017337005	249 51268 RNG RD 204	1965	

- We are going to focus on the population of “Residential” properties
- For simplicity, values are adjusted per $1000

```
#filter: residential houses
dat <- dat %>%
        filter(ASSESSCLAS=="Residential")  %>%
        mutate(assess_val = ASSESSMENT / 1000)
```

Population *vs.* sample

We can consider *all* residential properties in this county as a finite population or we can think that this is a random sample from an unknown data generating process.

To illustrate some concepts in this course we are going to consider this as a population and take a *random sample* of properties from it.

> **Heads up**: Recall the importance of using a *random* sample to obtain representative summaries and broad conclusions!

Important note

In general, we don't have information for all the population. We only have a *random sample* from the population and use it to *estimate* and *infer* information about the *unknown* population.

```
set.seed(561)

dat_s <- sample_n(dat, 1000, replace = FALSE)
```

Look at the top and the bottom of your data

```
dat_s  %>%  head(3)
dat_s  %>%  tail(3)
```

	the_geom	TAX_YEAR	ROLL_NUM	ADDRESS	YEAR_BUILT	AS
	<chr>	<int>	<dbl>	<chr>	<int>	
1	POINT (-113.26828568594681 53.53364376673626)	2015	-2147483648	107 GLENCOE BLVD	1990	
2	POINT (-113.31770257850987 53.516152401699564)	2015	-2147483648	116 WILLOW ST	1959	
3	POINT (-113.28777490571461 53.52360710210846)	2015	-2147483648	45 RAVEN DR	1966	

	the_geom	TAX_YEAR	ROLL_NUM	ADDRESS	YEAR_BUI
	<chr>	<int>	<dbl>	<chr>	<in
998	POINT (-113.29031567242677 53.55190602122994)	2015	-2147483648	106 COACHMAN WY	19
999	POINT (-113.28937772221407 53.515066319465284)	2015	-2147483648	45 MEADOWOOD CR	19
1000	POINT (-112.98115374382121 53.48441517042359)	2015	2114896000	254 52343 RNG RD 211	19

Check the packaging and check your “n”s

```
str(dat_s)
```

```
'data.frame':   1000 obs. of  17 variables:
 $ the_geom  : chr  "POINT (-113.26828568594681 53.53364376673626)" "POINT (-11
 $ TAX_YEAR  : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ ROLL_NUM  : num  -2.15e+09 -2.15e+09 -2.15e+09 -2.15e+09 1.22e+09 ...
 $ ADDRESS   : chr  "107 GLENCOE BLVD" "116 WILLOW ST" "45 RAVEN DR" "8 CRANBEI
 $ YEAR_BUILT: int  1990 1959 1966 2004 1982 2006 1993 1953 2008 2000 ...
 $ ASSESSCLAS: chr  "Residential" "Residential" "Residential" "Residential" ..
 $ BLDG_DESC : chr  "2 Storey & Basement" "1 Storey & Basement" "1 Storey & Bas
 $ BLDG_METRE: int  220 97 89 111 162 160 141 114 243 132 ...
 $ BLDG_FEET : int  2368 1044 960 1192 1740 1724 1515 1229 2613 1420 ...
 $ GARAGE    : chr  "Y" "Y" "N" "Y" ...
 $ FIREPLACE : chr  "Y" "Y" "N" "N" ...
 $ BASEMENT  : chr  "Y" "Y" "Y" "Y" ...
 $ BSMTDEVL  : chr  "Y" "Y" "Y" "Y" ...
 $ ASSESSMENT: int  536000 370000 318000 339000 543000 469000 388000 320000 11
 $ LATITUDE  : num  53.5 53.5 53.5 53.6 53.4 ...
 $ LONGITUDE : num  -113 -113 -113 -113 -113 ...
 $ assess_val: num  536 370 318 339 543 ...
```

Additional information about the data

We can define different random variables and study important qualities to characterize them and their relationships.

For example:

- the assessed value (the response Y in a LR model)
  - `assess_val` in the dataset
- the size (an explanatory variable X_1 in a LR model)
  - `BLDG_METRE` in the dataset
- the availability of a fireplace (another explanatory variable X_2 in a LR model)
  - `FIREPLACE` in the dataset

Get summary statistics

- Select the variables `assess_val` and `BLDG_METRE` and create a new dataframe called `dat_tax_SLR`
- Use the `gather` function to convert the dataset into a *long* format
- Then use `summarise` to obtain the maximum, minimum, median, mean, and standard deviation of these variables

> **Heads up**: for pedagogical purposes we start with this smaller dataset

```
# Your code goes here
dat_tax_SLR = dat_s |> select(assess_val, BLDG_METRE) |>
    gather() |>
    group_by(key) |>
    summarise(mean = mean(value, na.rm = TRUE),
              max = max(value),
              min = min(value),
              median = median(value, na.rm = TRUE),
              sd = sd(value, na.rm = TRUE))
dat_tax_SLR
```

A tibble: 2 × 6

key	mean	max	min	median	sd
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
BLDG_METRE	155.833	569	52	140	58.79801
assess_val	498.669	1784	107	453	188.72668

2.2 Population and sample distributions

As any random variable, these have a distribution function which we need to answer questions of the kind

> *what is the probability that the* `assess_val` *is greater than 500K*?

- Plot side-by-side histograms of the population and sample distributions of `assess_val`
- Add a vertical red line representing the position of the population and sample means (respectively)

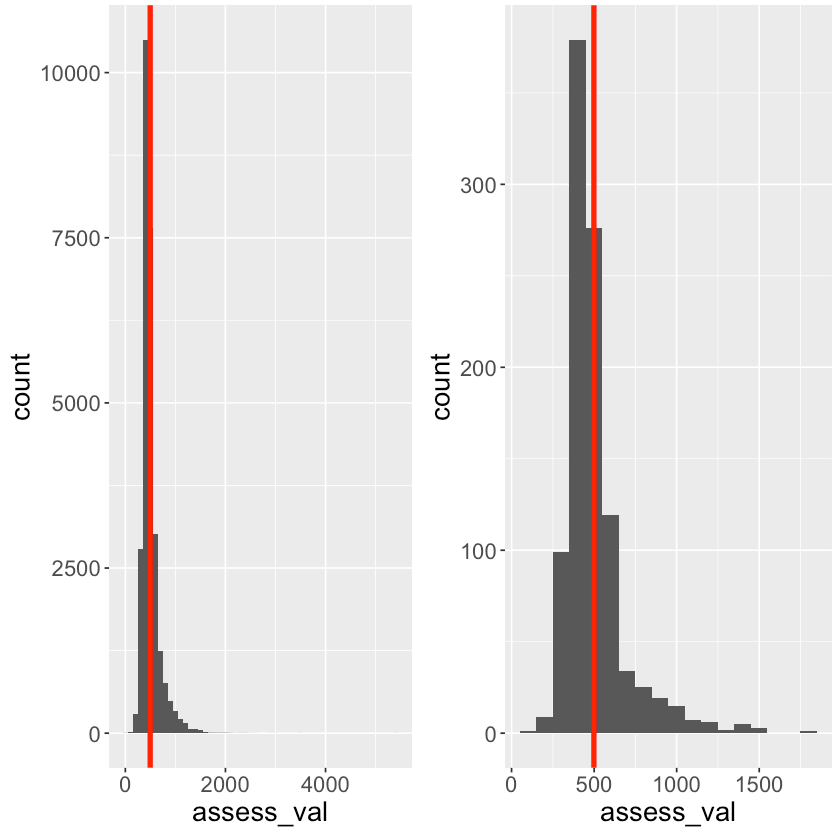
```

```

```r
# Your code goes here
library(cowplot)
# population
population <- dat |>
  ggplot(aes(x = assess_val)) +
  geom_histogram(binwidth=100) +
  theme(text = element_text(size = 16.5)) +
  geom_vline(xintercept = 496.4043, color = "red", size = 1.5)

# sample
sample <- dat_s |>
  ggplot(aes(x = assess_val)) +
  geom_histogram(binwidth=100) +
  theme(text = element_text(size = 16.5)) +
  geom_vline(xintercept = 498.669, color = "red", size = 1.5)

plot_grid(population, sample, nrow = 1)
```

```
Attaching package: ‘cowplot’
```

```
The following object is masked from ‘package:lubridate’:

    stamp
```

```
Warning message:
“Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.”
```



If we want to predict the assessed value of a property, the *best* guess would be its expected value $E[Y]$.

> Even if we don't have all the data from the population, we can make a good prediction if we know the average value

In our population data: $E[Y] = 496.4$

In our sample: $\bar{Y} = 498.669$

Important note

In general, we don't have information about the population and we use a *random sample* from the population to infer or predict

But if we want to predict the assessed value of a *large* property (e.g., more than 300 mts) ...

... it makes more sense to just focus on the values of large properties in our population

> Note: It is important that we only make predictions for our response variable within the observed range of our explanatory variable.

3.1 The conditional expectation

The *conditional* expected value is the *best* predictor given additional relevant information.

But, is the assessed value related to the size of the property? Is the "size" *relevant* information?

We can use a linear regression (LR) to answer these questions.

3.2 The regression line

Let $(X_i, Y_i) : i = 1, \ldots, n$ be a random sample of size n from the population

Then, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ E[\varepsilon_i | X_i] = E[\varepsilon_i] = 0$

- We made this assumption for any pair of random variables from our population!

Notation

- note the use of the subscript i to denote the ith experimental unit in our sample:
  - the ith house in our dataset
  - the ith patient in a medical study
  - the ith customer in an economics study

The error term, ε_i contains all factors affecting Y_i other than X_i

Assumptions

1. An important assumption is that the conditional expectation of the response is linearly related to the input variable and the line is the *linear regression*

$$E[\text{value}_i|\text{size}_i] = \beta_0 + \beta_1 \text{size}_i$$

or in a general context:

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

> **NOTE**: This is not the only way to model the conditional expectation! If the true conditional expectation is not linear, other methods will be better to predict the response! For example: in DSCI_571, you have used `kNN` (see lab problem)!

2. The random errors are independent and identically distributed: *iid assumption*
3. The random errors have all the same variance: $\text{Var}(\varepsilon_i) = \sigma^2$

Important note

As any other assumption, these may not hold or may not be a good assumption. We'll study later how to check these assumptions, some consequences of their violation and what to do in those cases.

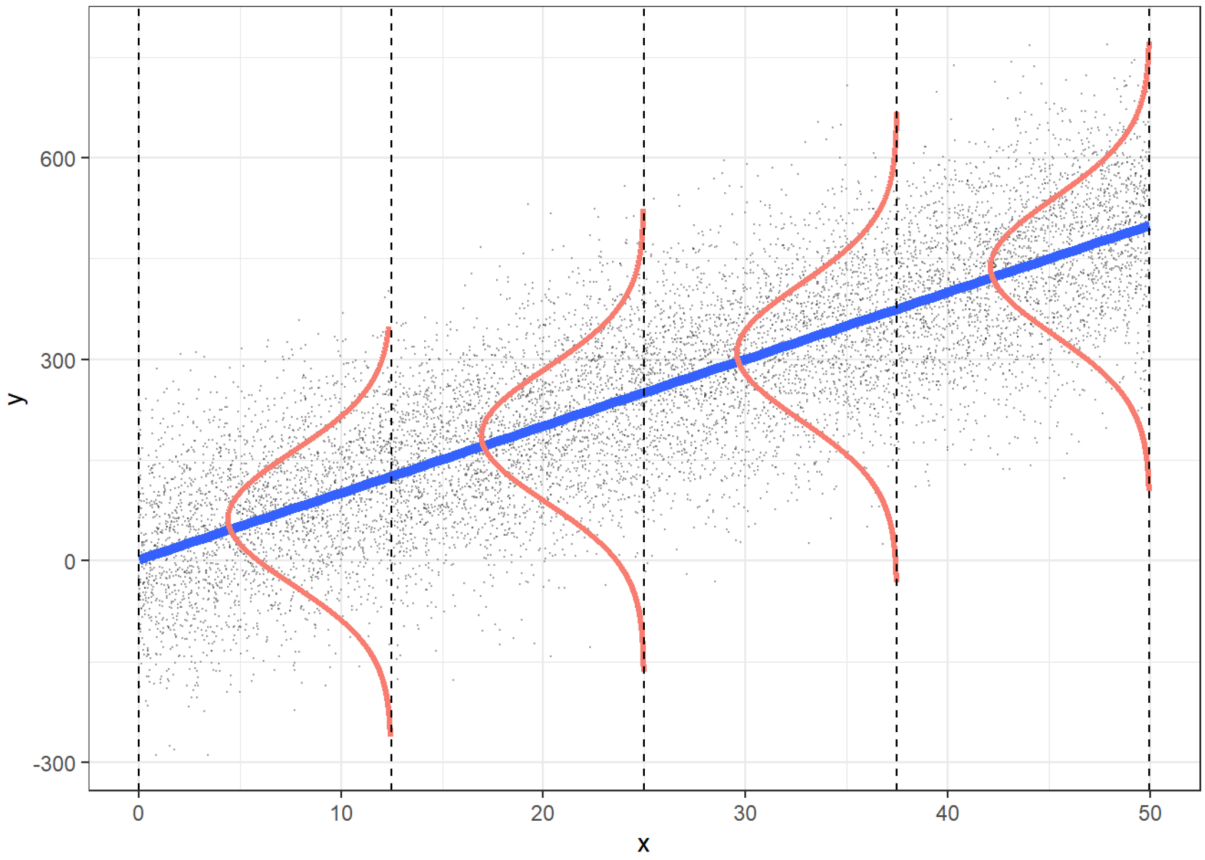


Figure 1.1: Assumptions for linear least squares regression (LLSR).

Image from the book: "Beyond Multiple Linear Regression", from Paul Roback and Julie Legler

https://bookdown.org/roback/bookdown-BeyondMLR/

3.3 Elements of a LR

The response variable: Y

Also known as: **output**, **explained variable**, **dependent variable**

The input variable(s): X

Also known as: **input**, **explanatory variables**, **independent variable**, **covariates**, **features**

> **Heads up**: In Simple Linear Regression (SLR) there's only 1 input variable

Question

You hypothesize that the cancer mortality of each county is related to the average level of poverty. Thus, you want to study quantitavely the association between `TARGET_deathRate` and `povertyPercent`.

Which variable would you use as the *response* and which as an *input* to build a LR?

A. Response: `TARGET_deathRate` ; Input: `povertyPercent`

B. Response: `povertyPercent` ; Input: `TARGET_deathRate`

The regression coefficients: $\beta_0,\ \beta_1$

The true intercept and the slope of this line are called *regression parameters or coefficients*

> Note: These population parameters are *unknown* and *non-random* (we use a *random sample* to estimate them using the `lm` function in R)

The error term: ε_i

The error term contains all factors affecting Y_i other than X_i

> Note: Any distributional assumption made about the error term also affect the random variable Y (if you assume that ε is a Normal random variable, then Y would also be Normal).

Question

True or false?

In a simple linear regression, the response is an exact linear function of the input variable, i.e.,

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

A. TRUE

B. FALSE

Question

True or false?

The population regression slope coefficient, β_1, is usually unknown so we need to estimate it using data.

A. TRUE

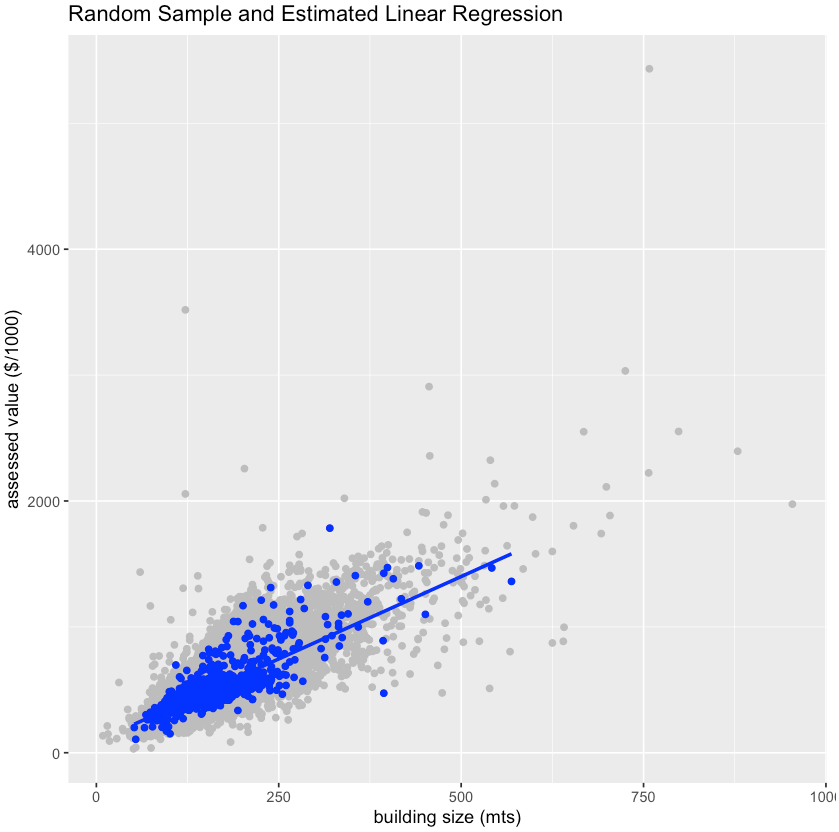
B. FALSE
The true regression parameters are unknown! So we use data from a random sample to estimate them!!

> in grey: all points in the population `dat`

> in blue: all points in the random sample `dat_s` (you can also use `dat_tax_SLR`)

```
plot_slr <- dat %>% ggplot(aes(BLDG_METRE, assess_val)) +
  geom_point(color = "grey") +
  xlab("building size (mts)") +
  ylab("assessed value ($/1000)") +
  geom_point(data = dat_s, aes(BLDG_METRE, assess_val), color = "blue") +
  geom_smooth(method = lm, se = FALSE, color = "blue", data=dat_s) +
  ggtitle("Random Sample and Estimated Linear Regression")

plot_slr
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Question

How would you choose the **best line**?

A. The line that contains most data points

B. The line that minimizes the distance of the points to the line

C. The line that looks the best upon visual inspection

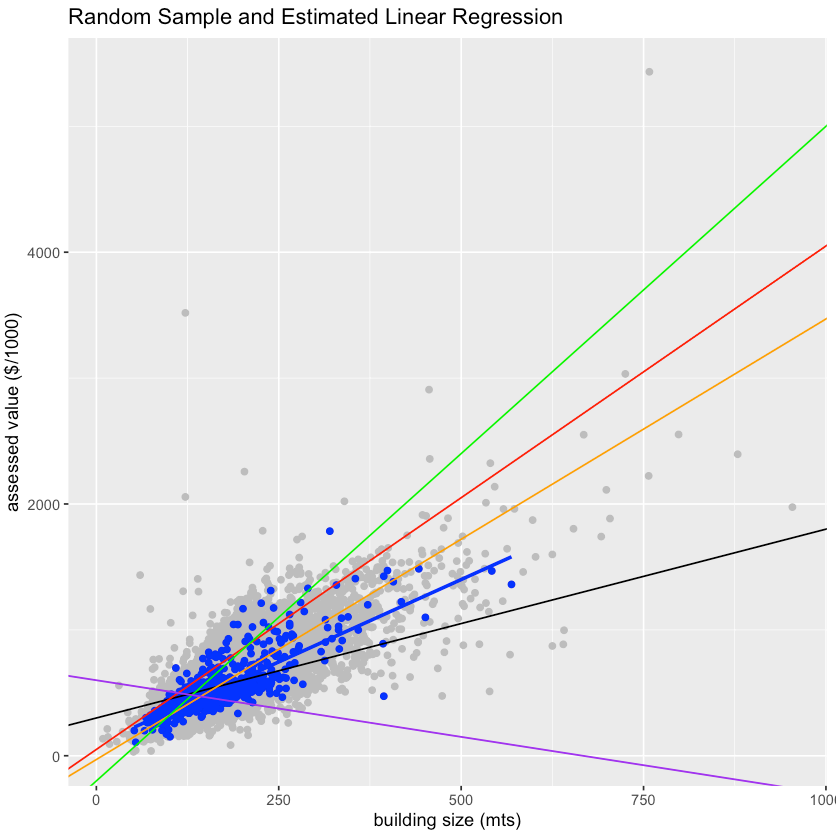
4.1 Least squares estimation

How do we find the *best* line?

- This is not the only line we can fit!

```
plot_slr  +
    geom_abline(intercept=300,slope=1.5, col = "black")+
    geom_abline(intercept=-30,slope=3.5, col = "orange")+
    geom_abline(intercept=50,slope=4, col = "red")+
    geom_abline(intercept=-200,slope=5.2, col = "green")+
    geom_abline(intercept=600,slope=-0.9, col = "purple")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Discussion :

To define the **best line** we need to know how to measure the distance of the points to the line!

Which of the following criteria would you choose to define “distance of a point to the line”?

> **Note**: discuss with your choice with your neighbour, more than one option can be possible

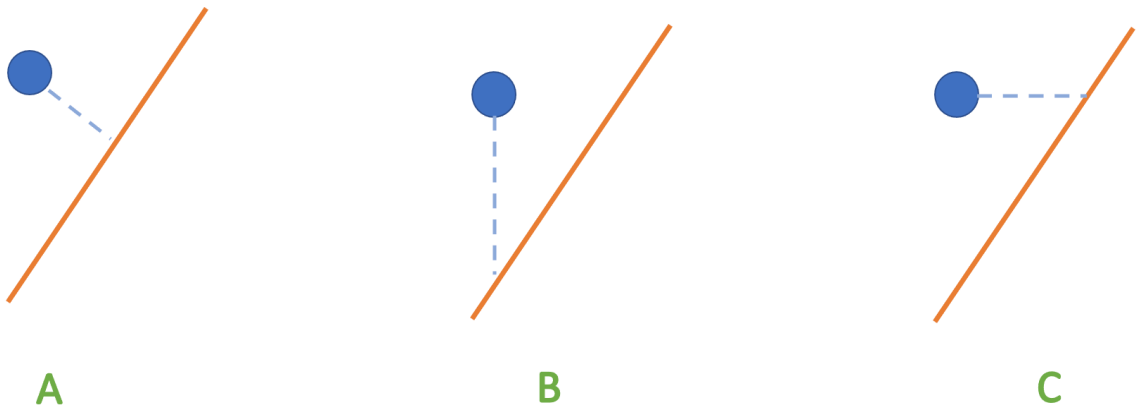


Figure by Prof. Joel Ostblom

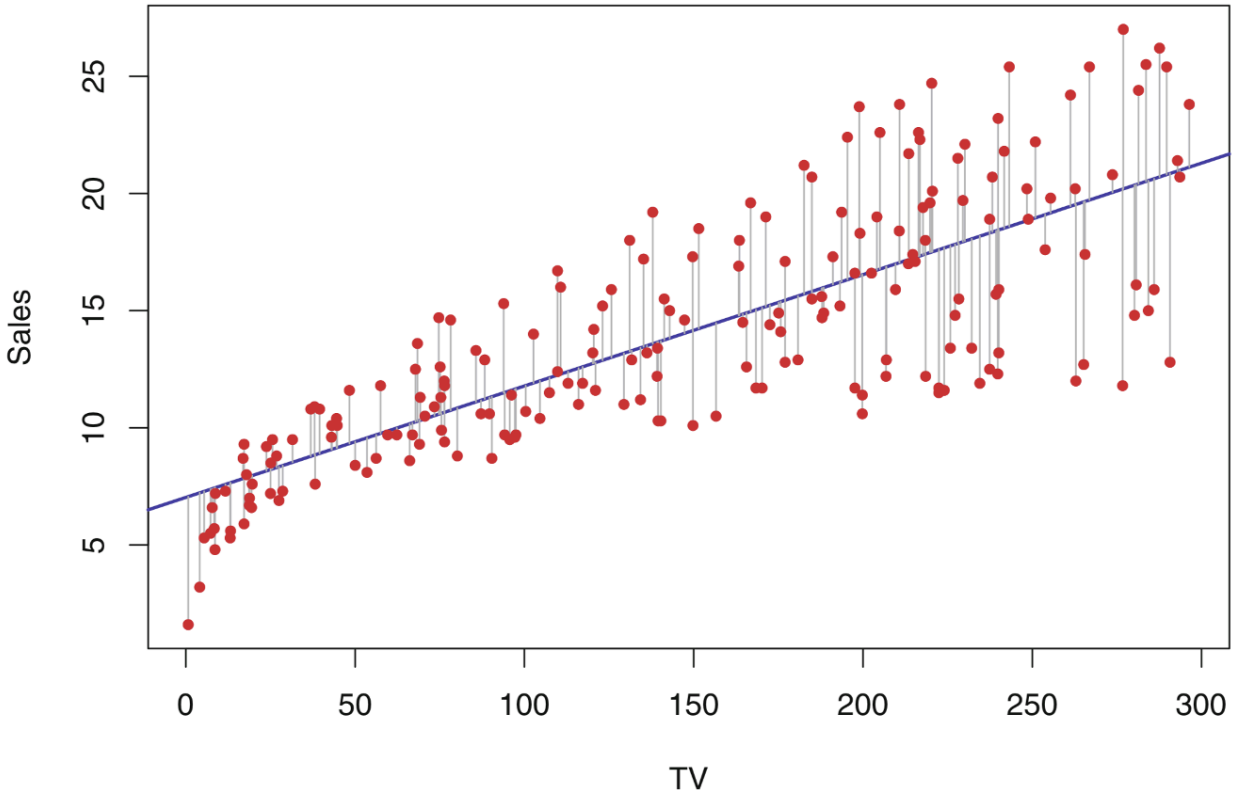


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression*

From An Introduction to Statistical Learning, by James, Witten, Hastie and Tibshirani

Least Squares method minimizes the sum of the squares of the residuals. The residuals are the difference between the observed value of the response (y_i) and the predicted value of the response ($\hat{y}_i$):

$$r_i = y_i - \hat{y}_i$$

> **Heads up:** the residuals are the vertical distances of each point to the estimated line

Check this application

> This is just one criteria to estimate the regression line

4.2 LS in R

The parameters of the linear model, a.k.a. regression coefficients, can be estimated through **ordinary least squares**! We will use the `lm` function in R to obtain estimates using data.

> **Heads up**: LS is not the only method to estimate the regression coefficients. However, it is the default method in `lm`

The relevant arguments are:

- `formula`: takes the form `response ~ input`.
- `data`: takes a data frame in tidy format.

> **Note**: `lm(response ~ .,data= df)` uses all variables in the dataset `df`, except the `response`, as predictors

> **Note**: `lm(response ~ input - 1,data= df)` forces the estimated intercept to be 0. Never do this unless you know what you are doing and why.

```
lm_slr <- lm(assess_val ~ BLDG_METRE, data = dat_s)
tidy(lm_slr) %>%
    mutate_if(is.numeric, round, 3)
```

A tibble: 2 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	90.769	9.793	9.268	0
BLDG_METRE	2.618	0.059	44.514	0

4.3 Parameter *vs* Estimator *vs* Estimate

3 important different concepts:

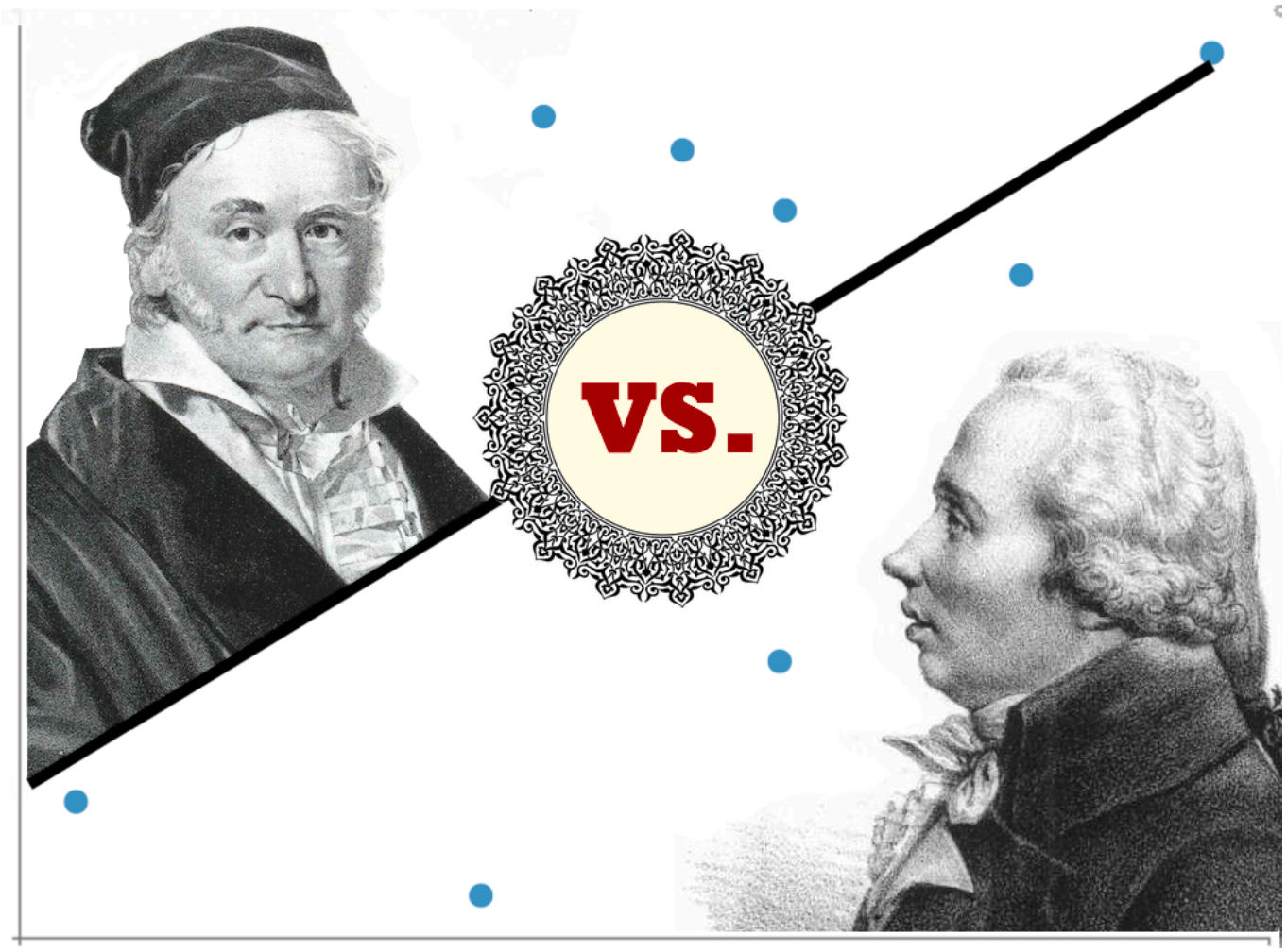
Course	Population Parameter	Estimator	Estimate
	unknown quantity	function of the random sample: *random variable*	real number computed with data (non-random)
DSCI552	mean = μ = $E[Y]$	sample mean = $\bar{Y}$	516.6
DSCI561	slope = β_1	estimator of the slope = $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$	2.7

> **Note**: usually $\hat{\beta}_0$ and $\hat{\beta}_1$ are used for both the estimates and the estimators, which can be confusing

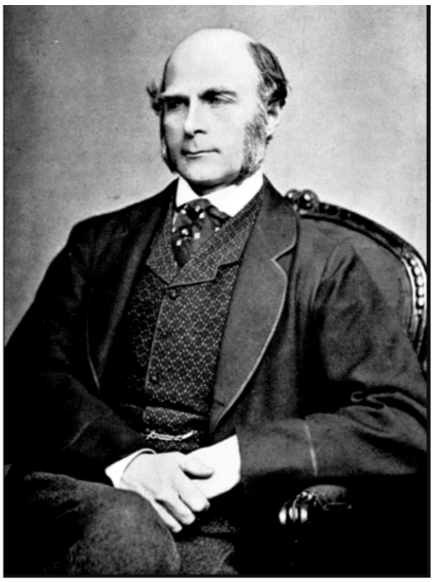
Historical throwback ...

> **Historical note I**: Least squares (a classical method in Regression) was first used by **Legendre** (1805) and by **Gauss** (1809) to estimate the orbits of comets based on measurements of the comets' previous locations. Gauss even predicted the appearence

of the asteroid Ceres using LS combined with other complex computations (Source: The Discovery of Statistical Regression)



Historical note II: However, neither Legendre or Gauss coined the term "Regression". **Francis Galton** in the nineteenth century use this term to describe a biological phenomenon that he observed: "It appeared from these experiments that the offspring did not tend to resemble their parents seeds in size, but to be always more mediocre than they". It was later his colleage **Karl Pearson** who associated Least Squares to Regression...



246 *Anthropological Miscellanea.*

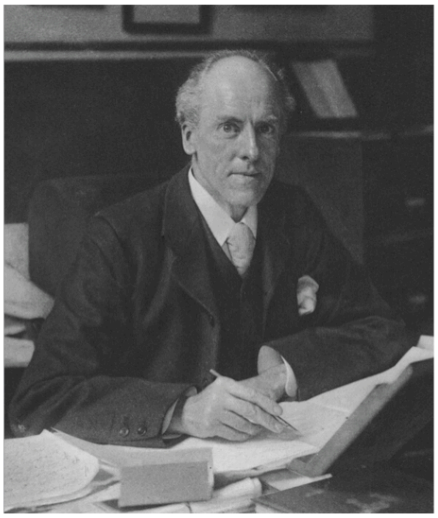
ANTHROPOLOGICAL MISCELLANEA.

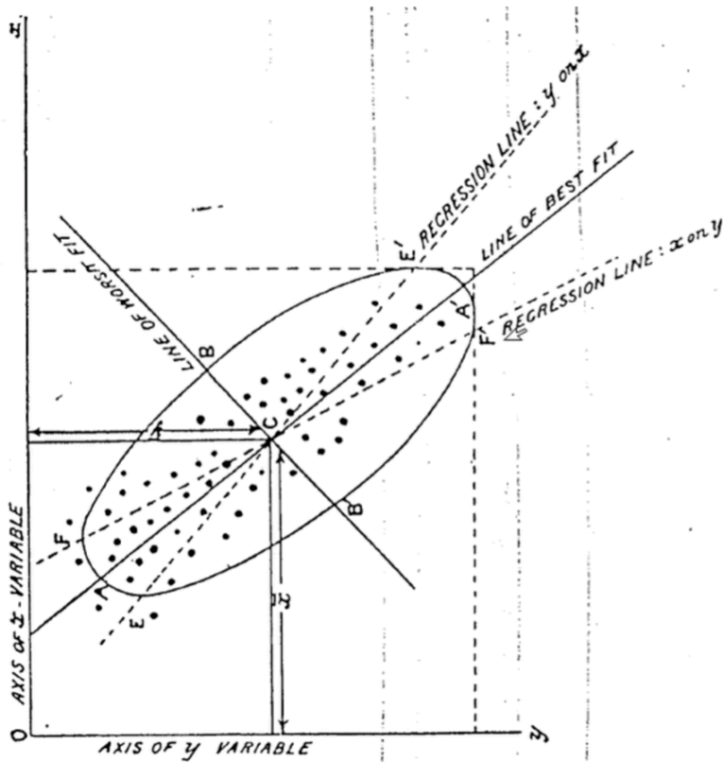
Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address





Note: unfortunately, Francis Galton had disturbing and unacceptable views of race (https://en.wikipedia.org/wiki/Francis_Galton)