Statistical Literacy — MINT

Lecture 10: Correlation & Linear Univariate Regression

Rémi Viné

The Graduate Institute | Geneva

November 27th, 2023

Outline

Housekeeping

- Problem set 10 is now available
- **Exam**:
  - ⇒ You can use all material put on Moodle (for this course) and your notes
  - ⇒ The use of internet (apart from Moodle itself) and any means of communication are strictly forbidden
  - ⇒ Recall: need of a calculator, a laptop (with an access to Moodle), some pens (we provide draft)

Link among variables

- Hypothesis testing for means: quantitative variables (in this course)
  - *E.g.*, mean battery life in minutes in two situations: compare the sample mean *versus* the producer's statement
- Link between two variables: $Y \Leftrightarrow X$
- Correlation: compare two quantitative variables
- Regression: compare two (or more) quantitative variables
  - Note that some variable may not be quantitative
  - Y is the **explained** variable (or dependent, or endogenous)
  - X is the **explanatory** variable (or independent, or exogenous)
- **Beware: showing a correlation is not showing a causality**

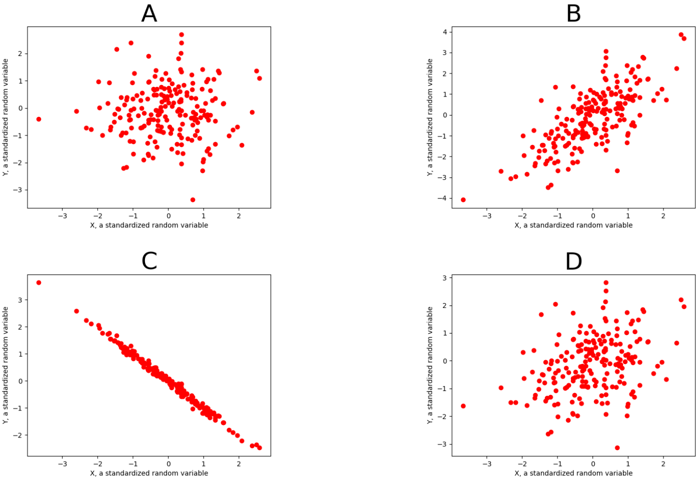
What is a correlation?

Focus on linear correlation

- **Covariance**: how do two variables vary together (**covary**)?
  - In sample: $Cov(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
  - Pay attention at the unit of measurement
- **Correlation**: scale the covariance
  - In population: $Correlation = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$
  - In sample: $Correlation = r_{XY} = \frac{Cov(X,Y)}{s_X s_Y}$
    - $-1 \leq r_{XY} \leq +1$
    - This is about **Linear** correlation
    - Interpret: (i) Sign, (ii) Strength, (iii) Linearity

Linear correlation between numerical variables

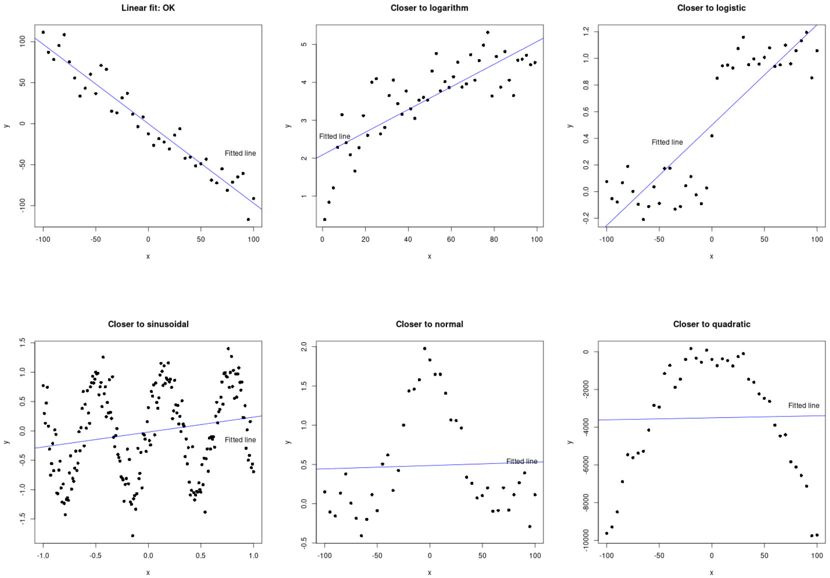
Use scatterplots to show the relationships of 2 numerical variables



- What's what?

  1. $r = +0.349$; 2. $r = +0.036$ 3. $r = +0.723$; 4. $r = -0.995$

Lapalissade: Linear correlation is about **linear** correlation

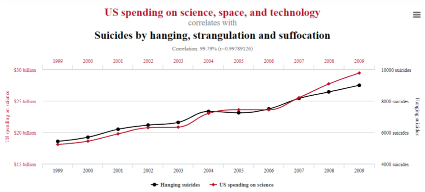


Spurious correlations

Random possibility and/or confounding factors

- A strong correlation is not enough, some common sense is required!

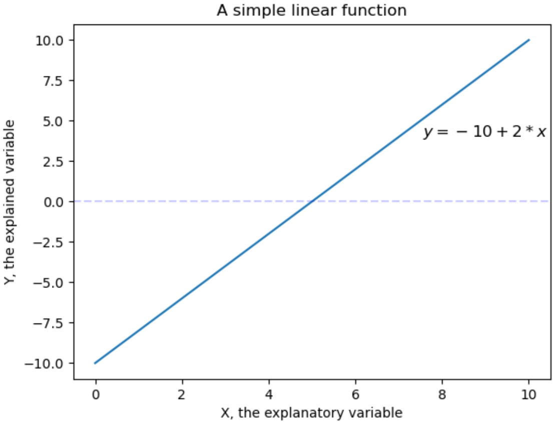
(a) Random correlation (Source)



(b) Confounding factor (Source)

A step beyond correlation: the univariate regression

- Remember the linear function?
  - $y = f(x) = b_0 + b_1 x$

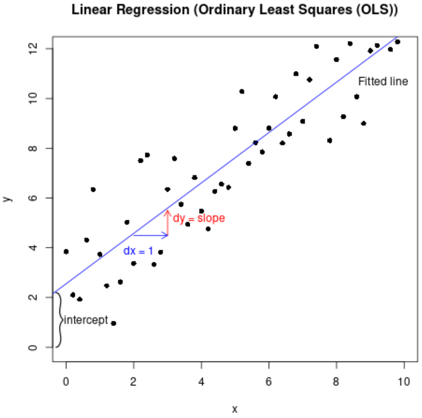


A univariate regression is a linear function summarizing the data

Sample regression

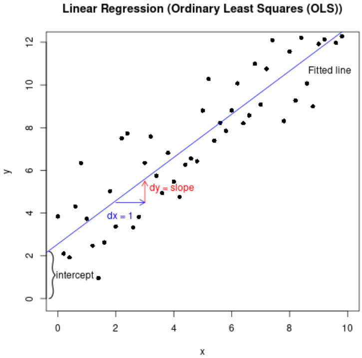
- $\hat{y} = b_0 + b_1 x$
  - The "hat" notation is about the **predicted value**, *i.e.*, the line

- The gaps between the line and the data points are **errors**: $y = b_0 + b_1 x + e$
- Errors: $e_i = y_i - \hat{y}_i$
  - where "i" stands for any observation in the data



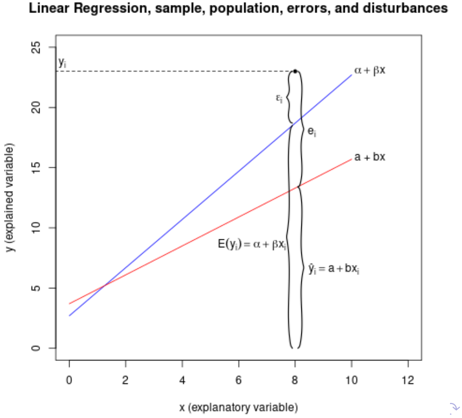
Interpret the coefficients

- $\hat{y} = b_0 + b_1 x$
  - b_0 is the intercept
    - The predicted value of y when $x_1 = 0$
    - Not always meaningful, context-dependent (is $x = 0$ realistic?)
  - b_1 is the slope
    - What is the corresponding change of y when x change by $+1$



Univariate regression for **inference**: from the sample regression to the population regression

- Sample regression $y = b_0 + b_1 x + e$
- Population regression $y = \beta_0 + \beta_1 x + \varepsilon$
  - With ε the disturbances
- The hope is that the sample and the population regressions are close enough (so that $b_0 \approx \beta_0$ and $b_1 \approx \beta_1$)

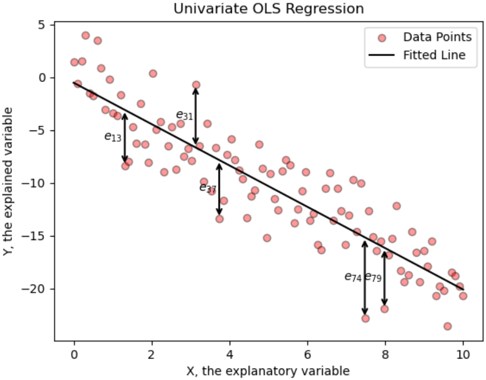


But how is this "fitted" line obtained?

The Ordinary Least Squares (OLS)

- This line is the result of a $Minimization$ procedure
  - Find the coefficients b_0 & b_1 $Minimize$ the **sum of squared errors**

$$\rightarrow \; Min_{b_0,b_1} \sum_i e_i^2 = Min_{b_0,b_1} \sum_i (y_i - \hat{y}_i)^2$$



- For univariate regression (no need to remember this by heart)

$$\Rightarrow \; b_0 = \bar{y} - b_1 \bar{x}$$

$$\Rightarrow \; b_1 = \frac{Cov(X,Y)}{Var(X)}$$

Example: look at data

Source: Our World in Data

Data excerpt:

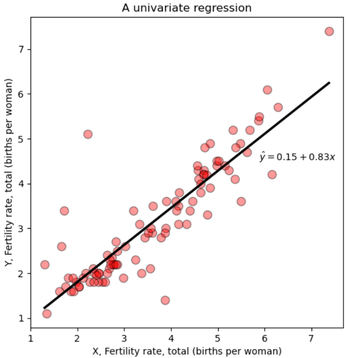
	Entity	Code	Year	Wanted fertility rate (births per woman)	Fertility rate, total (births per woman)	Population (historical estimates)	Continent
1	Afghanistan	AFG	2015	4.4	4.976	33753500.0	Asia
598	Albania	ALB	2018	1.6	1.617	2877019.0	Europe
1487	Angola	AGO	2016	5.2	5.686	29154742.0	Africa
2392	Armenia	ARM	2016	1.7	1.744	2865841.0	Asia
3609	Azerbaijan	AZE	2006	1.8	1.970	8763353.0	Asia
4393	Bangladesh	BGD	2018	1.7	2.036	163683952.0	Asia
5682	Benin	BEN	2018	4.9	4.836	11940688.0	Africa
6288	Bolivia	BOL	2008	2.0	3.364	9880593.0	South America
6874	Botswana	BWA	1988	3.9	4.839	1261276.0	Africa
7136	Brazil	BRA	1996	1.8	2.536	166037120.0	South America

- Assume the interest is on the link between the wanted fertility (Y) and the actual fertility (X)
  - Here, each “i” is a country

Example: build a regression line

Source: Our World in Data

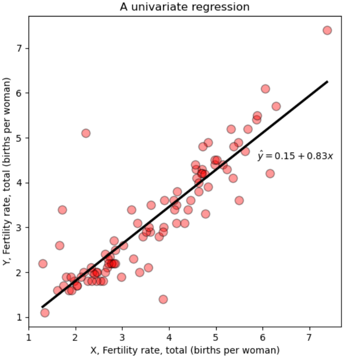
	Explained Variable:
	Wanted fertility
(standard errors in parentheses)	(1)
Actual Fertility rate (b_1)	0.8253***
	(0.0460)
const (b_0)	0.1541
	(0.1791)
Observations	94
R^2	0.7774
Residual Std. Error	0.6072
	(df = 92)
F Statistic	321.3239***
	(df = 1.0; 92.0)
Note:	*p<0.1; **p<0.05; ***p<0.01



Example: interpretation

Source: Our World in Data

	Explained Variable:
	Wanted fertility
(standard errors in parentheses)	(1)
Actual Fertility rate (b_1)	0.8253***
	(0.0460)
const (b_0)	0.1541
	(0.1791)
Observations	94
R^2	0.7774
Residual Std. Error	0.6072
	(df = 92)
F Statistic	321.3239***
	(df = 1.0; 92.0)
Note:	*p<0.1; **p<0.05; ***p<0.01

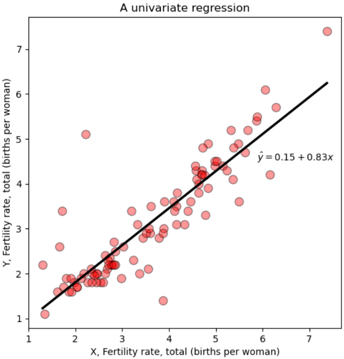


- Intercept: in a country with zero fertility, the corresponding wanted fertility would be 0.15 child (not very meaningful in this context)
- In a given country, if the actual fertility were higher by one child, the corresponding wanted children number would be higher by 0.83

Example: prediction

Source: Our World in Data

	Explained Variable:
	Wanted fertility
(standard errors in parentheses)	(1)
Actual Fertility rate (b_1)	0.8253***
	(0.0460)
const (b_0)	0.1541
	(0.1791)
Observations	94
R^2	0.7774
Residual Std. Error	0.6072
	(df = 92)
F Statistic	321.3239***
	(df = 1.0; 92.0)
Note:	*p<0.1; **p<0.05; ***p<0.01



- For 5 actual children per woman, what is the predicted number of wanted children?
  - $\hat{y} = 0.1541 + 0.8253 * 5 = 4.2806$ children

Inference in regression

- Using a sample, the idea is to generalize to all the population of interest
  - Here, all the countries in the world (assuming no selection bias)
- From b_1, what could be β_1?
  - If it is different than 0, then there is a statistically significant relationship between X and Y
- *Note that a similar procedure can be done for the intercept (but this is less interesting even in a univariate case)*

Inference in regression

- Run a hypothesis testing (check that $n > 30$ for validity)

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

  - Note that one can proceed to a one tail test as seen in last lecture, and change the tested value (instead of 0)

- Test-statistic (univariate) $= \frac{estimated\ coefficient - hypothesized\ coefficient}{standard\ error} = \frac{b_1 - 0}{\frac{s}{\sqrt{n}}}$
  - Or look at the $p - value$

- Critical value: $t^{n-2}_{\alpha/2}$ *(beware of the degrees of freedom)*
  - Look at the Student table (unless you have the population variance - *very unlikely*)

Inference in regression

- Compare
  - Test-statistic *versus* critical value
  - *p-value versus* α
- Reject H_0 if
  - $|Test - statistic| \geq |critical\ value|$
  - *p-value* $\leq \alpha$
- If H_0 is not rejected, one concludes that there is no significant relationship between $X\&Y$ at a level of significance α
- Alternatively: one concludes that X is not a statistically significant (at a level of significance α) predictor for Y in this model
  - And opposite interpretation if H_0 is **NOT** rejected

Example: inference

Test whether X is a significant predictor at $\alpha = 1\%$

	Explained Variable:
	Wanted fertility
(standard errors in parentheses)	(1)
Actual Fertility rate (b_1)	0.8253***
	(0.0460)
const (b_0)	0.1541
	(0.1791)
Observations	94
R^2	0.7774
Residual Std. Error	0.6072
	(df = 92)
F Statistic	321.3239***
	(df = 1.0; 92.0)
Note:	*p<0.1; **p<0.05; ***p<0.01

- Reading the table, the $p - value$ of b_1 is lower than 0.01, the α (looking at "stars")

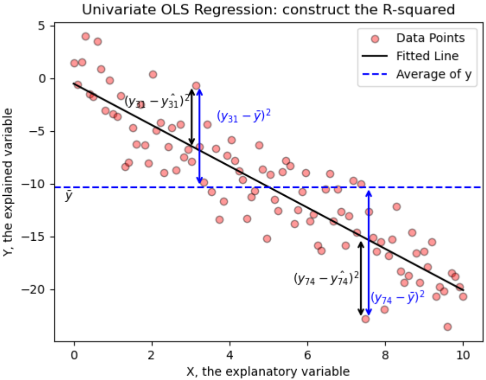
$\Rightarrow$ H_0 is rejected, the actual fertility rate is a significant predictor of the wanted fertility rate at $\alpha = 1\%$

- Alternatively, one could compute the test-statistic: $\frac{0.8253-0}{0.0460} \approx 17.9413$
  - Looking at the Student table $t^{92}_{0.005} \approx t^{100}_{0.005} = 2.626$
  - Hence, $|test - statistic| > |critical\ value|$

How good is the model?

The fit of the model

- The **R-squared** (or R^2): the share of the variance in Y that is captured by the model compared to a model without X



- $R^2 = \frac{ExplainedVariance}{TotalVariance} = 1 - \frac{ResidualVariance}{TotalVariance} = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$
- $0 \leq R^2 \leq 1$
- The closer to 1, the better the fit of the model
- In univariate case: $r_{XY} \times r_{XY} = R^2$

Example: How good is the model?

	Explained Variable:
	Wanted fertility
(standard errors in parentheses)	(1)
Actual Fertility rate (b_1)	0.8253***
	(0.0460)
const (b_0)	0.1541
	(0.1791)
Observations	94
R^2	0.7774
Residual Std. Error	0.6072
	(df = 92)
F Statistic	321.3239***
	(df = 1.0; 92.0)
Note:	*p<0.1; **p<0.05; ***p<0.01

- $R^2 = 0.7774$

$\Rightarrow$ 77.74% of the variance of the wanted fertility is captured by the univariate regression used

- The correlation coefficient can be retrieved: $r_{XY} = (Sign(b_1) \times \sqrt{R^2}) = +\sqrt{0.7774} \approx 0.8817$

- The correlation coefficient between the actual fertility and the wanted fertility is positive, strong and linear, as the correlation coefficient is $+0.8817$

Inference: is the model useful in predicting Y?

Comparing our model with a model without explanatory variables: the $F-test$

- The model brings predictive power as long as as least one explanatory variable is statistically significant
  - This test is mostly useful for **multivariate** regressions (next week)
- The **F-test** for regression, step by step:
  1. Validity: Normal Sampling distribution
  2. Hypotheses:
  $$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H_a : \ At\ least\ one\ \beta_i \neq 0 \end{cases}$$
  3. Test-statistic $= \frac{Explained\ Variance}{Unexplained\ Variance}$
  4. Critical value: $F_{\alpha}^{(k);(n-k-1)}$
  5. Compare test-statistic & Critical value OR $p-value$ & α

Example: The F-test for regression

	Explained Variable:
	Wanted fertility
(standard errors in parentheses)	(1)
Actual Fertility rate (b_1)	0.8253***
	(0.0460)
const (b_0)	0.1541
	(0.1791)
Observations	94
R^2	0.7774
Residual Std. Error	0.6072
	(df = 92)
F Statistic	321.3239***
	(df = 1.0; 92.0)
Note:	*p<0.1; **p<0.05; ***p<0.01

- $test - statistic = 321.3229$
  - $F_{0.01}^{1;92} = 6.919$ (Source)
- $p - value < 0.01$ (as shown by the "stars")

$\Rightarrow$ At $\alpha \geq 1\%$, H_0 is rejected so that the model is statistically useful in predicting the wanted fertility

- Here again, the **F-test** is not so useful in univariate (a test the explanatory variable's coefficient would suffice), but it will be useful when different explanatory variables will be included in the model

Quick summary on univariate OLS

1. What is the OLS method?
2. How to interpret the coefficients?
3. How to make predictions?
4. How to make coefficient-wise inference?
5. What is the overall fit of the model?
   - 5.1 Magnitude: the $R - squared$
   - 5.2 Statistical significance: the $F - test$

- One (big) missing element $\rightarrow$ why is OLS method used this much?

Next session

- Next session is on multivariate linear regression and OLS assumptions.
  - It will be our last lecture with new content
  - Lecture 12 will be a wrap-up session