Applied Statistical Analysis I

Bivariate regression, inference & prediction

Trajche Panov, PhD
panovt@tcd.ie

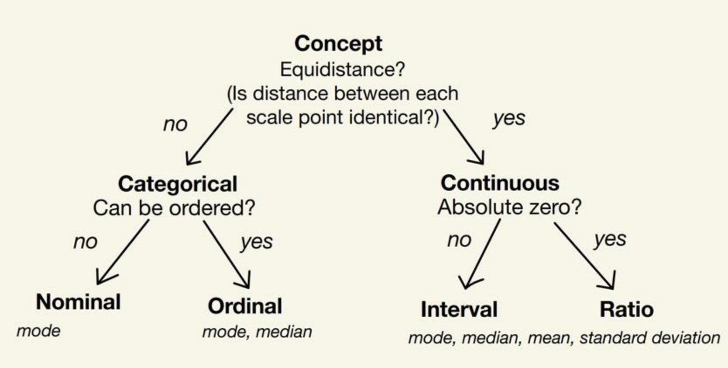
Department of Political Science
Trinity College Dublin

October 01, 2023

Today's Agenda

(1) Lecture recap
(2) Git pull
(3) Tutorial exercises

Importance of measurement scales



(Kellstedt and Whitten 2018, Chap. 5)

<u>Discrete:</u> finite set of possible values (Contingency tables, chi-square test)
<u>Continuous:</u> infinite set of possible values (t-test for mean and difference in means, correlation, scatter plot, dependent variable in linear regression)

Correlation

What is correlation? How can we measure correlation?

How can we test the statistical significance of correlation?

Correlation

What is correlation?

- "The *correlation* between two features of the world is the extent to which they tend to occur together" (Bueno de Mesquita and Fowler 2021, 13).
- "If two features of the world tend to occur together, they are *positively correlated*" (Bueno de Mesquita and Fowler 2021, 13).
- "If the occurrence of another feature of the world is unrelated to the occurrence of another feature of the world, they are *uncorrelated*" (Bueno de Mesquita and Fowler 2021, 13).
- "And if when one feature of the world occurs the other tends not to occur, they are *negatively correlated*" (Bueno de Mesquita and Fowler 2021, 13).

Correlation

How can we measure correlation?

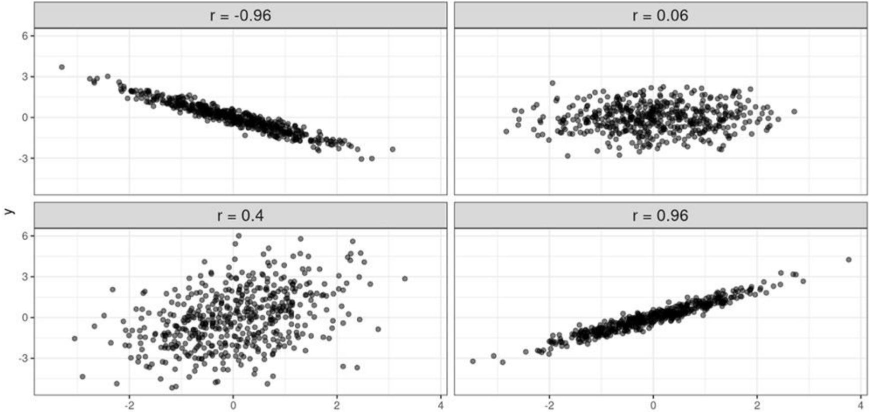
- Covariance: covariance is the average of the product of deviations of two quantitative variables from the mean, $cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$
- Positive association, if larger-than-average X_i co-occurs with larger-than-average Y_i, and vice versa.
- Negative association, if larger-than-average X_i co-occurs with smaller-than-average Y_i, and vice versa.
- only interpret sign, not magnitude of association, given that covariance is scale-dependent

Correlation

How can we measure correlation?

- Correlation: (correlation coefficient, Pearson correlation coefficient, Pearson's r, r) standardized average of the product of deviations of two variables from the mean (=standardized covariance)
- **standardize covariance through dividing by product of standard deviations of the two variables,** $r_{XY} = \frac{covariance(XY)}{S_X S_Y}$
- ranges between -1 and 1, with 0=no association, the larger the absolute value, the stronger the association

Correlation



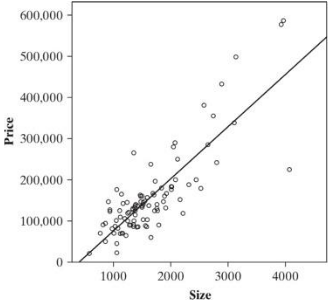
Correlation

How can we test the statistical significance of correlation?

- Null and alternative hypotheses:
  - there is no association between X and Y, $\rho_{xy} = 0$ (H_0)
  - there is an association between X and Y, $\rho_{xy} /= 0$ (H_a)
- Test statistic: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (in R)
- Test statistic: $t = \frac{r}{\sqrt{1-r^2/n-2}}$ (in Agresti 2018)

Correlation

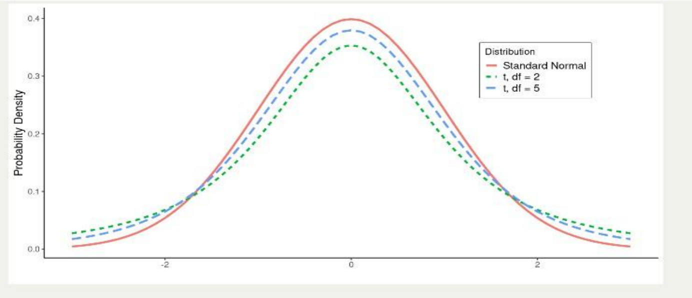
How can we test the statistical significance of correlation?



- Is there an association between house selling price and size (Agresti 2018, 278–283)? $r = 0.83378$
- $t = \sqrt{\frac{r}{1-r^2/n-2}} = \sqrt{\frac{0.834}{(1-0.695)/98}} = 14.95$
- How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that H_0 is true? → Probability distribution

Correlation

How can we test the statistical significance of correlation?



What is the conclusion? P-value < 0.05, We can reject H_0 with an error probability (p-value) of essentially 0% (p=0.0001). → There is an association between house selling price and size

Shortcomings of correlation analysis

- no indication on the “substantive importance or size of the relationship between X and Y” (Bueno de Mesquita and Fowler 2021, 29).
- Slope: “tells us, descriptively, how much Y changes, on average, as X increases by one unit” (Bueno de Mesquita and Fowler 2021, 29).

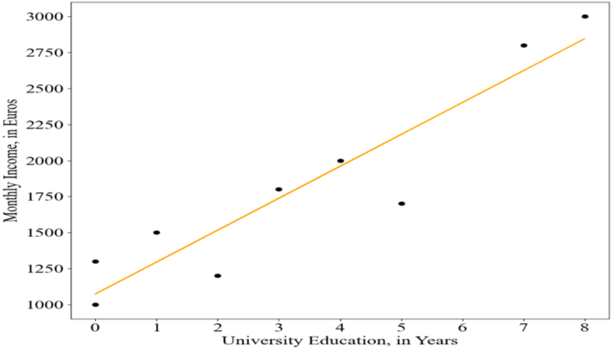
Linear regression model

What is a linear regression model? What interpretations can we make?

Regression analysis

What is a linear regression model?

- Find linear line of best fit, $Y_i = \alpha + \beta X_i + \epsilon_i$



Regression analysis

What is a linear regression model?

- Find linear line of best fit, $Y_i = \alpha + \beta X_i + \epsilon_i$
- α (intercept): expected value of Y when $X = 0$
- β (slope): expected change in Y when X increases by one unit
- $\hat{Y}$ (expected value): predicted outcome based on the regression model, $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
- ϵ (error/residual): difference between actual and predicted outcome, $\epsilon_i = Y_i - \hat{Y}_i$

Example – Canada 2005

According to human capital theory, increased education is associated with greater earnings.

Random sample of 22 Saskatchewan males aged 35-39 with positive wages and salaries in 2004, from the Survey of Labour and Income Dynamics, 2005.

Let *x* be total number of years of school completed (YRSCHL18) and *y* be wages and salaries in dollars (WGSAL42).

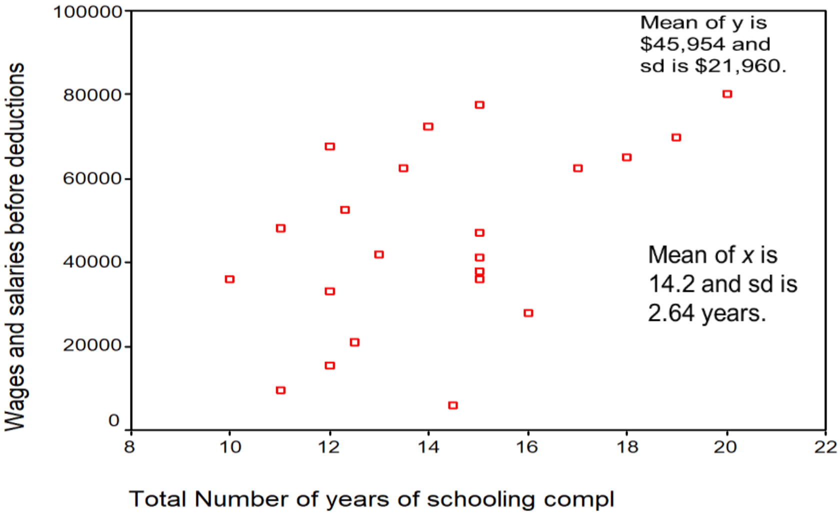
Source: Statistics Canada, Survey of Labour and Income Dynamics, 2005 [Canada]: External Cross-sectional Economic Person File [machine readable data file]. From IDLS through UR Data Library.

ID#	YRSCHL18	WGSAL42
1	17	62500
2	12	15500
3	12	67500
4	11	9500
5	15	38000
6	15	36000
7	19	70000
8	15	47000
9	20	80000
10	16	28000
11	18	65000
12	11	48000
13	14	72500
14	12	33000
15	14.5	6000
16	13.5	62500
17	15	77500
18	13	42000
19	10	36000
20	12.5	21000
21	15	41000
22	12.3	52500

YRSCHL18 is the variable “number of years of schooling”

WGSAL42 is the variable “wages and salaries in dollars, 2004”

Plot of WGSAL42 with YRSCHL18



Analysis of the Results

H_0: $\beta_1 = 0$. Schooling has no effect on earnings.
H_1: $\beta_1 > 0$. Schooling has a positive effect on earnings.

From the least squares estimates, using the data for the 22 cases, the regression equation and associate statistics are:

$y = -13{,}493 + 4{,}181\ x$.
$R^2 = 0.253$, $r = 0.\ 503$.
Standard error of the slope b_0 is 1,606.
$t = 2.603$ (20 df), significance = 0.017.
At $\alpha = 0.05$, reject H_0, accept H_1 and conclude that schooling has a positive effect on earnings.

Each extra year of schooling adds $4,181 to annual wages and salaries for those in this sample.

Expected wages and salaries for those with 20 years of schooling is -13,493 + (4,181 x 20) = $70,127.

Equation of a line

$y = \beta_0 + \beta_1 x$. x is the independent variable (on horizontal) and y is the dependent variable (on vertical).

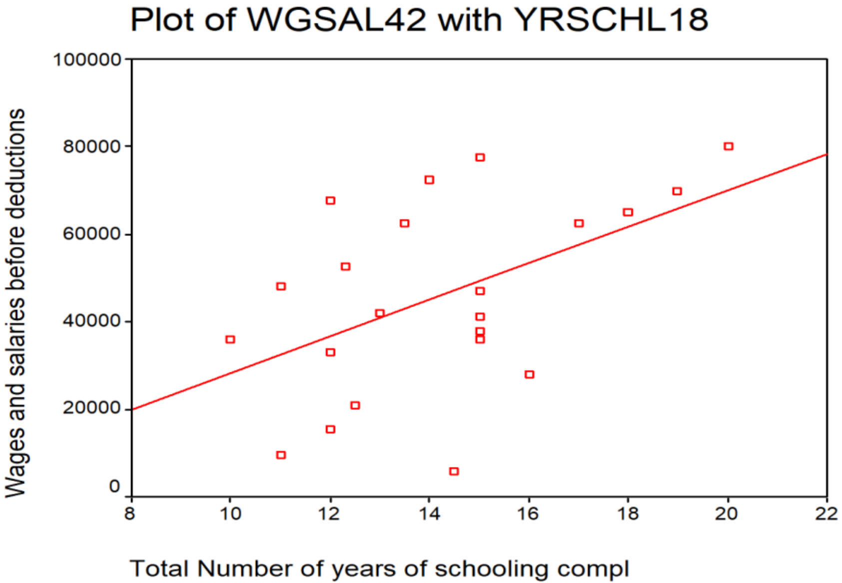
β_0 and β_1 are the two parameters that determine the equation of the line.

β_0 is the y intercept – determines the height of the line.

β_1 is the slope of the line.

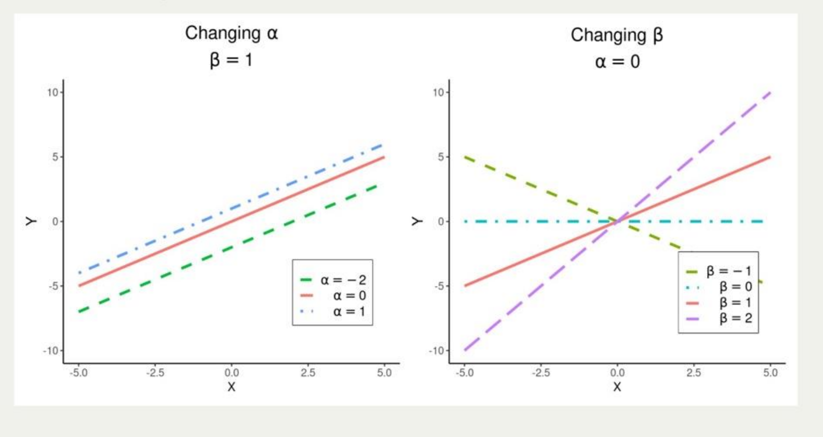
Positive, negative, or zero.

Size of β_1 provides an estimate of the manner that x is related to y.


Plot of WGSAL42 with YRSCHL18
100000
80000
60000
40000
20000
0
Wages and salaries before deductions
8
10
12
14
16
18
20
22
Total Number of years of schooling compl

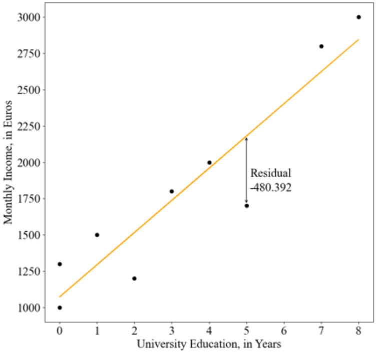
Regression analysis

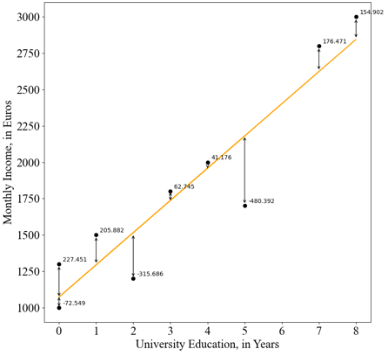
Varieties of linear relationships



Regression analysis

What interpretations can we make? (residuals)





OLS Line

The least squares line is the unique line for which the sum of the squares of the deviations of the y values from the line is as small as possible.
Minimize the sum of the squares of the errors ε.
Or, equivalent to this, minimize the sum of the squares of the differences of the y values from the values of E(y). That is, find b0 and b1 that minimize

$$\sum \varepsilon^2 = \sum \left(y_i - \hat{y}_i\right)^2 = \sum \left(y_i - b_0 - b_1 x\right)^2$$

Province	Income	Alcohol
Newfoundland	26.8	8.7
Prince Edward Island	27.1	8.4
Nova Scotia	29.5	8.8
New Brunswick	28.4	7.6
Quebec	30.8	8.9
Ontario	36.4	10
Manitoba	30.4	9.7
Saskatchewan	29.8	8.9
Alberta	35.1	11.1
British Columbia	32.5	10.9

Is alcohol a superior good?

Income is family income in thousands of dollars per capita, 1986. (independent variable)

Alcohol is litres of alcohol consumed per person 15 years of age or over, 1985-86. (dependent variable)

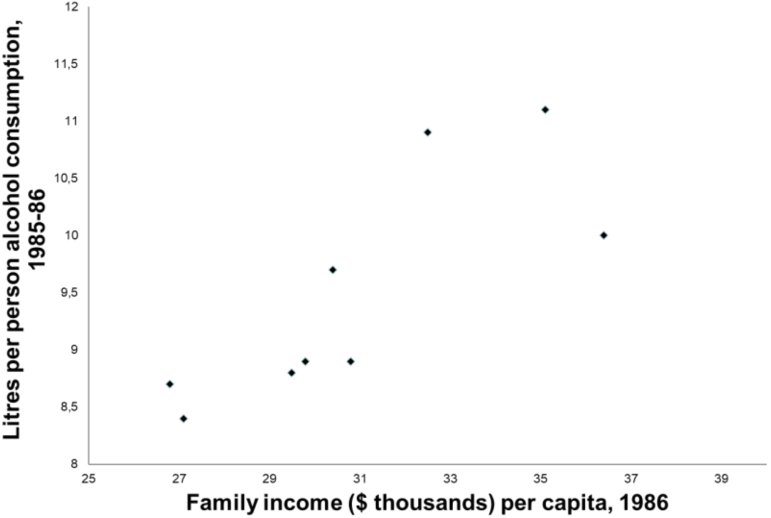
Sources: Saskatchewan Alcohol and Drug Abuse Commission, *Fast Factsheet*, Regina, 1988
Statistics Canada, Economlc Families – 1986 [machine-readable data file, 1988.

Hypotheses

H_0: $\beta_1 = 0$. Income has no effect on alcohol consumption.

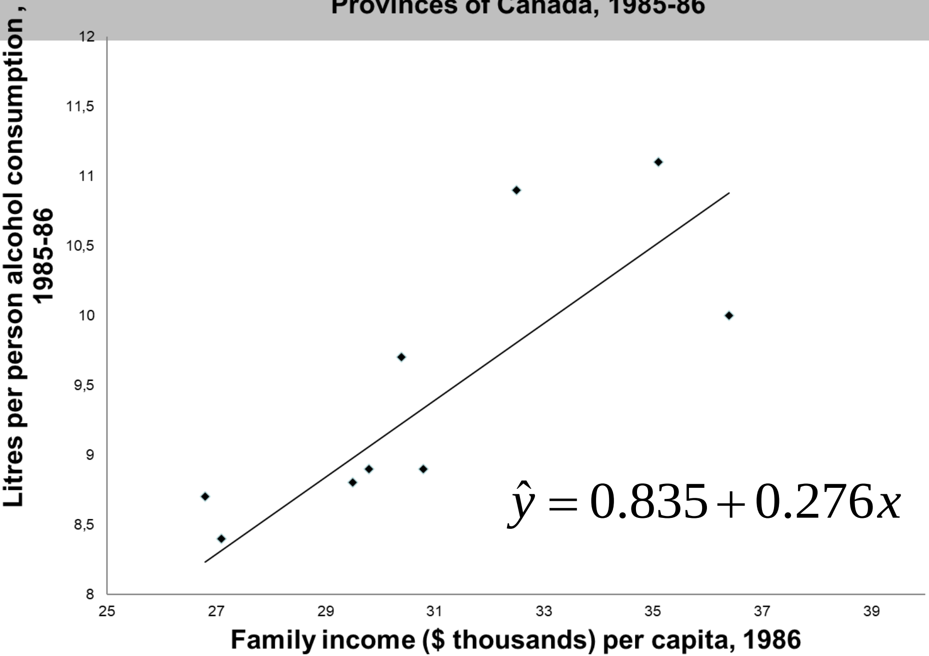
H_1: $\beta_1 > 0$. Income has a positive effect on alcohol consumption.

Scatter diagram of alcohol consumption (y) with income (x), Provinces of Canada, 1985-86



Province	x	y	x-barx	y-bary	(x-barx)(y-bary)	x-barx sq
Newfoundland	26.8	8.7	-3.88	-0.6	2.328	15.0544
PEI	27.1	8.4	-3.58	-0.9	3.222	12.8164
Nova Scotia	29.5	8.8	-1.18	-0.5	0.59	1.3924
New Brunswick	28.4	7.6	-2.28	-1.7	3.876	5.1984
Quebec	30.8	8.9	0.12	-0.4	-0.048	0.0144
Ontario	36.4	10	5.72	0.7	4.004	32.7184
Manitoba	30.4	9.7	-0.28	0.4	-0.112	0.0784
Saskatchewan	29.8	8.9	-0.88	-0.4	0.352	0.7744
Alberta	35.1	11.1	4.42	1.8	7.956	19.5364
British Columbia	32.5	10.9	1.82	1.6	2.912	3.3124
sum	306.8	93	-6.8E-14	-7.1E-15	25.08	90.896
mean	30.68	9.3				
				b1	0.275919732	
				b0	0.834782609	

Scatter diagram of alcohol consumption (*y*) with income (*x*), Provinces of Canada, 1985-86



$$\hat{y} = 0.835 + 0.276x$$

0 00 0 0 000 00

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.790288
R Square	0.624555
Adjusted R Square	0.577624
Standard Error	0.721104
Observations	10

Analysis. b_1 = 0.276 and its standard error is 0.076, for a *t* value of 3.648. At α = 0.01, the null hypothesis can be rejected (ie. with H_0, the probability of a *t* this large or larger is 0.0065) and the alternative hypothesis accepted. At 0.01 significance, there is evidence that alcohol is a superior good, ie. that income has a positive effect on alcohol consumption.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.920067	6.920067	13.30803	0.006513
Residual	8	4.159933	0.519992		
Total	9	11.08			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.834783	2.331675	0.358018	0.729592
X Variable 1	0.27592	0.075636	3.648018	0.006513

Uses of regression line

- **Draw line** – select two x values (eg. 26 and 36) and compute the predicted y values (8.1 and 10.8, respectively). Plot these points and draw line.

$$\hat{y} = 0.835 + 0.276x = 0.832 + (0.276 \times 26) = 8.091$$

$$\hat{y} = 0.835 + 0.276x = 0.832 + (0.276 \times 36) = 10.771$$

- **Interpolation**. If a city had a mean income of $32,000, the expected level of alcohol consumption would be 9.7 litres per capita.

$$\hat{y} = 0.835 + 0.276x = 0.832 + (0.276 \times 32) = 9.667$$

Goodness of Fit

- *y* is the dependent variable, or the variable to be explained.
- How much of *y* is explained statistically from the regression model, in this case the line?
- Total variation in *y* is termed the total sum of squares, or SST.

$$\mathrm{SST} = \sum (y_i - \bar{y})^2$$

- The common measure of goodness of fit of the line is the **coefficient of determination**, the proportion of the variation or SST that is "explained" by the line.

Interpretation of R^2

Proportion, or percentage if multiplied by 100, of the variation in the dependent variable that is statistically explained by the regression line.

$0 \leq R^2 \leq 1$.

Large R^2 means the line fits the observed points well and the line explains a lot of the variation in the dependent variable, at least in statistical terms.

Small R^2 means the line does not fit the observed points very well and the line does not explain much of the variation in the dependent variable.

- Random or error component dominates.
- Missing variables.
- Relationship between x and y may not be linear.

How large is a large R^2?

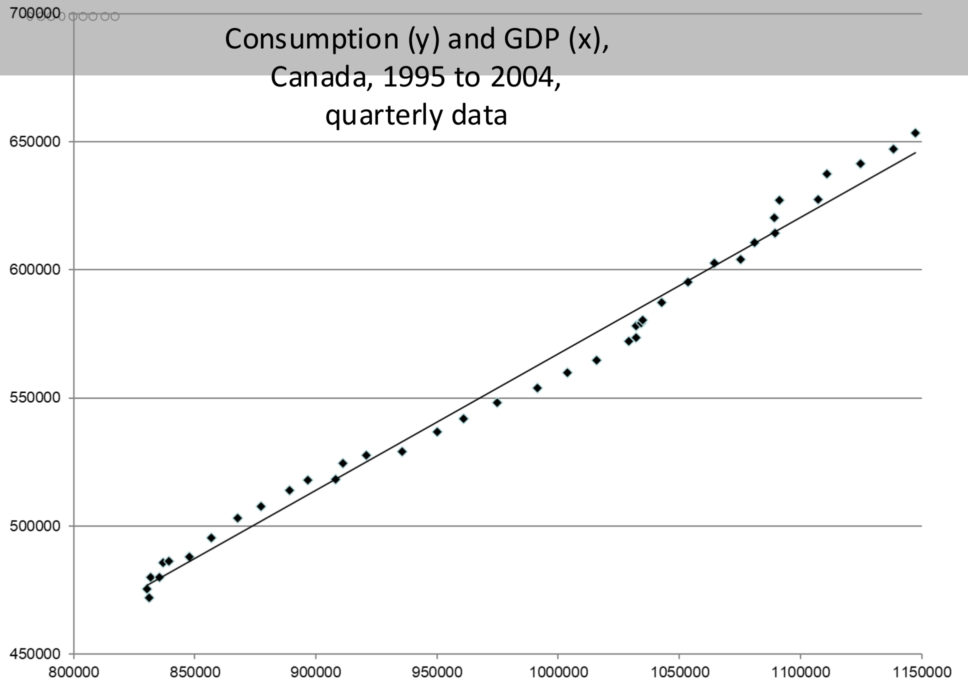
Extent of relationship – weak relationship associated with low value and strong relationship associated with large value.

Type of data

Micro/survey data associated with small values of R^2. For schooling/earnings example, $R^2 = 0.253$. Much individual variation.

Grouped data associated with larger values of R^2. In income/alcohol example, $R^2 = 0.625$. Grouping averages out individual variation.

Time series data often results in very high R^2. In consumption function example (next slide), $R^2 = 0.988$. Trends often move together.


Consumption (y) and GDP (x), Canada, 1995 to 2004, quarterly data
700000
650000
600000
550000
500000
450000
800000
850000
900000
950000
1000000
1050000
1100000
1150000

Beware of R^2

Difficult to compare across equations, especially with different types of data and forms of relationships.

More variables added to model can increase R^2.

Adjusted R^2 can correct for this. Grouped or averaged observations can result in larger values of R^2.

Need to test for statistical significance.

We want good estimates of β_0 and β_1, rather than high R^2.

At the same time, for similar types of data and issues, a model with a larger value of R^2 may be preferable to one with a smaller value.

Ordinary least squares (OLS)

How are intercept and slope estimated?

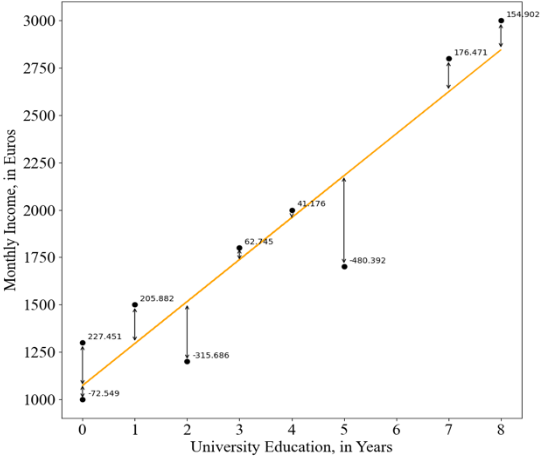
Ordinary least squares (OLS)

How are intercept and slope estimated?

- How do we find the line which best fits the data?
- Apply the OLS (Ordinary Least Squares) method, which minimizes the sum of squared errors (SSE).
- Sum of squared errors = the sum of squared differences between actual and predicted values of Y.
- $SSE = \sum_{i=1}^{n}(\hat{\epsilon_i})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\alpha} - \hat{\beta}X_i))^2$
  $\rightarrow$ minimize this!

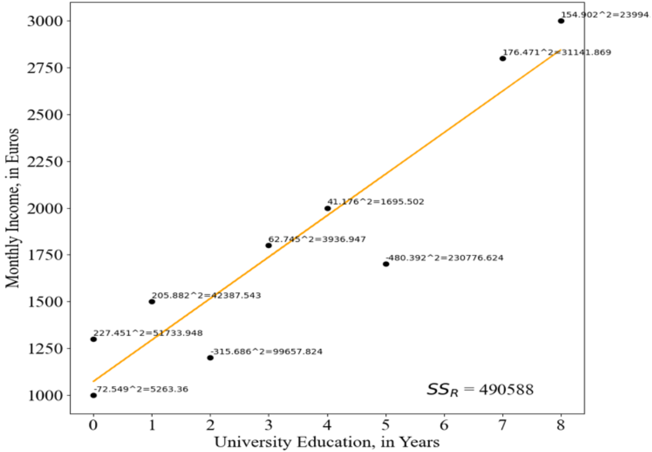
Ordinary least squares (OLS)

How are intercept and slope estimated?



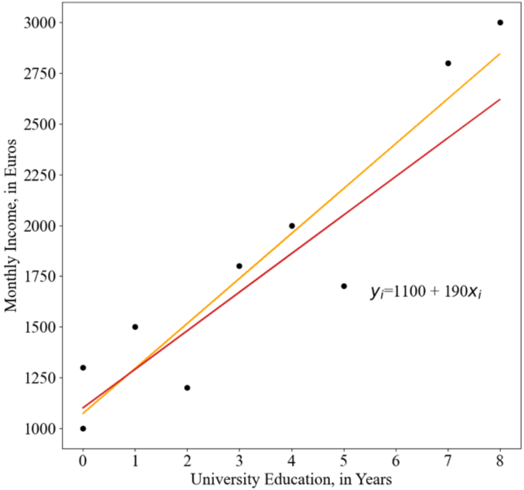
Ordinary least squares (OLS)

How are intercept and slope estimated?



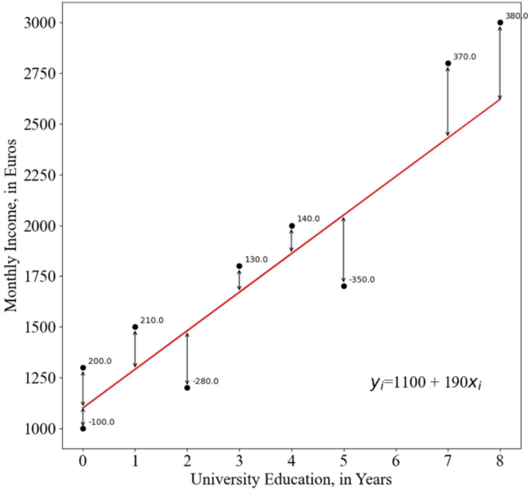
Ordinary least squares (OLS)

How are intercept and slope estimated?



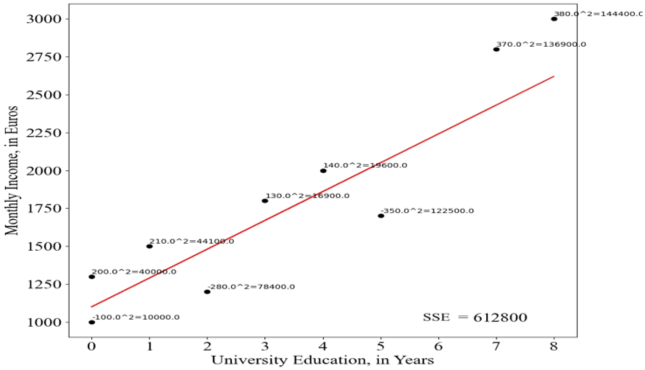
Ordinary least squares (OLS)

How are intercept and slope estimated?



Ordinary least squares (OLS)

How are intercept and slope estimated?



$612,800 > 490,588 \dashrightarrow SSE_{(RED)} > SSE_{(ORANGE)}$

$\dashrightarrow$ Orange regression line has better fit.

Inference about the slope

What is the t-test for the slope of a regression line?

Inference about the slope

What is the t-test for the slope of a regression line?

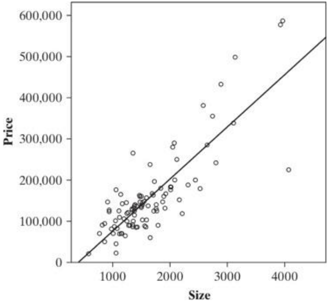
- Null and alternative hypotheses:
  - there is no association between X and Y, $\beta = 0$ (H_0)
  - there is an association between X and Y, $\beta \neq 0$ (H_a)
- Test statistic: "measures the number of standard errors between the estimate and the H_0 value" (Agresti 2018, 192).

$$t = \frac{\text{Estimate of parameter} - \text{Null hypothesis value of parameter}}{\text{Standard error of estimate}}$$

$$t = \frac{\hat{\beta} - \beta_{H_0}}{se_{\hat{\beta}}} = \frac{\hat{\beta}}{se_{\hat{\beta}}},\ H_0 \text{ assumes } \beta = 0$$

Inference about the slope

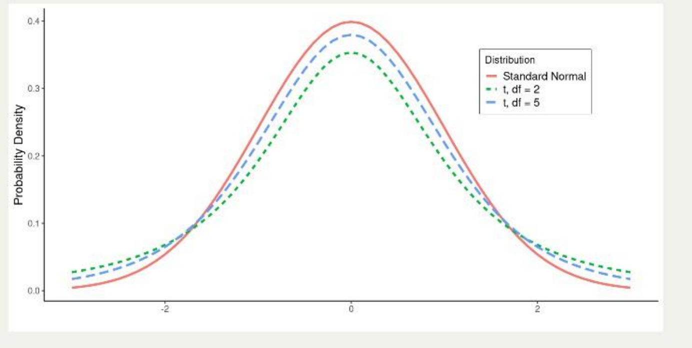
What is the t-test for the slope of a regression line?



- Is there an association between house selling price and size (Agresti 2018, 278–280)? $Price = 50,926.2 + 126.6 * Size$
- $t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} = \frac{126.6}{8.47} = 14.95$
- How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that H_0 is true? $\rightarrow$ Probability distribution

Inference about the slope

What is the t-test for the slope of a regression line?

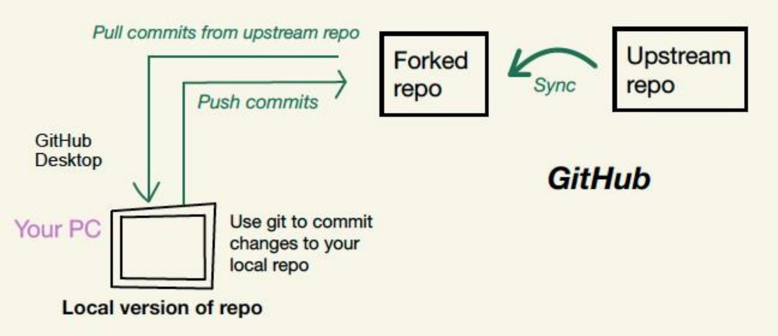


What is the conclusion? P-value < 0.05, We can reject H_0 with an error probability (p-value) of essentially 0% (< 0.0001). → There is an association between house selling price and size

Software check

How to update your local repository? How to git pull?

Software check



1. Synchronize fork
2. Fetch origin

References I

Agresti, Alan. 2018. *Statistical methods for the social sciences.* Harlow: Pearson.

Bueno de Mesquita, Ethan, and Anthony Fowler. 2021. *Thinking clearly with data: A guide to quantitative reasoning and analysis.* Princeton: Princeton University Press.

Kellstedt, Paul M., and Guy D. Whitten. 2018. *The fundamentals of political science research.* Cambridge: Cambridge University Press.