Simple Regression: OLS, Interpretation, ANOVA, and R-squared

Lecture 5

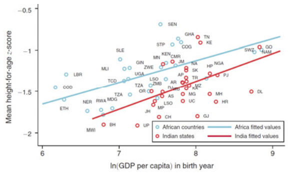
Reading: Sections 7.1 – 7.7, 19.4

Why Are Indian Children So Short?

- “One in four children under age five, worldwide, is so short as to be stunted (UNICEF 2014). Child stunting—a key marker of child malnutrition—[can mean adults that] are less healthy, have lower cognitive ability, and earn less.”
- “[India’s] child stunting rate is over 40 percent, an outlier even among poor countries (IIPS 2010).”
- “[We use] survey data on over 168,000 children from India and 25 African countries.”

Excerpts, pp. 2600-01 from “Why Are Indian Children So Short? The Role of Birth Order and Son Preference” (2017) https://doi.org/10.1257/aer.20151282

Child Height versus National GDP



- “Figure 1 graphs average child height-for-age for sub-Saharan African countries [blue] and Indian states [red] against income.”
- “Both regions exhibit a positive correlation between income and child height, but the curve for India [red] is lower; at a given level of income, Indian children are shorter.”
- “Given that India performs better than African countries on most health and development indicators, this contrast is striking and is the focus of this paper.”

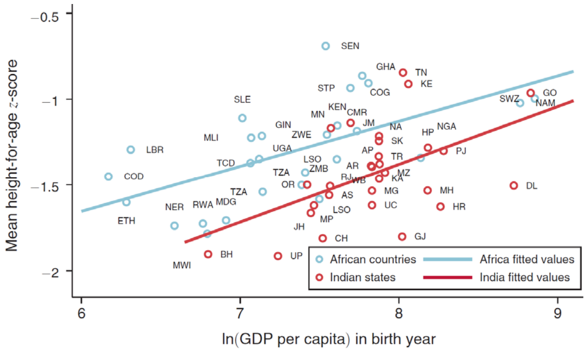


Figure 1. Child Height versus National GDP

Notes: The light and dark circles represent sub-Saharan African countries and Indian states, respectively. The averages are calculated over all children less than 60 months old. The lines represent the best linear fit for each sample. National GDP data are based on the Penn World Table 9.0 (Feenstra, Inklaar, and Timmer 2015).

HFA z-scores [Excerpts, pp. 2604-05]

“[We] create the child’s height-for-age (HFA) z-score based on the World Health Organization (WHO) universally applicable growth standard for children aged zero to five years.[10]

A z-score of 0 represents the median of the gender- and age-specific reference population, and a z-score of -2 indicates that the child is two standard deviations below that reference-population median, which is the cutoff for being stunted.

Our primary outcome of interest is the HFA z-score because it is the child health measure that has been most often linked to later-life outcomes and is viewed as the best cumulative measure of child malnutrition.”

[10] The WHO standard describes how children should grow if they receive proper nutrition and health care. It is premised on the fact that the height distribution among children under age five who receive adequate nutrition and health care has been shown to be similar in most ethnic groups (de Onis et al. 2006; WHO Multicentre Growth Reference Study Group 2006a).

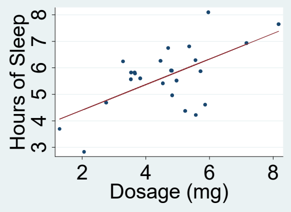
Regression Analysis: Least Squares Method

- Ordinary Least Squares (OLS): Quantitative method to fit a line through a scatter diagram
- Formula for a geometric line: $y = a + bx$
  - a is the y-intercept and b is the slope
  - x variable, independent variable, RHS variable
  - y variable, dependent variable, LHS variable
  - But careful, interpreting an OLS line is *not* the same as for a geometric line
- Goal: Find line that “best” fits data (a and b)

Results of a Double-Blind Drug Trial

i	Dosage (mg) x_i	Sleep (hrs) y_i
1	5.9	4.6
2	3.5	5.8
3	7.2	6.9
4	3.6	5.8
…	…	…
25	8.2	7.6

X-bar = 4.61 Y-bar = 5.66
s_x = 1.51 s_y = 1.18
s_{xy} = 1.09



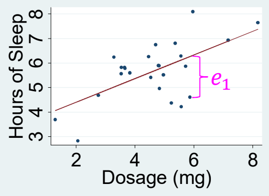
Next, discuss how to find the red OLS line

OLS: Returns the a (intercept) and b (slope) that minimize SSE

- Predicted value of y (y-hat): $\hat{y}_i = a + bx_i$
- Residual: $e_i = y_i - \hat{y}_i$
- OLS minimizes the SSE:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\boxed{b = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}} \quad \boxed{a = \bar{y} - b\bar{x}}$$



$$b = \frac{s_{xy}}{s_x^2} = \frac{1.09}{1.51^2} = 0.48$$

$$a = \bar{y} - b\bar{x} = 5.66 - 0.48 * 4.61 = 3.44$$

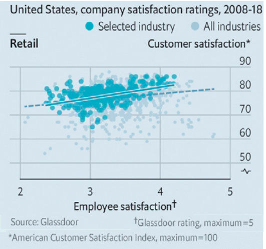
$$\hat{y}_i = 3.44 + 0.48x_i$$

Interpreting an OLS line

- Unlike a regular line, remember most (or all) of the data does not lie on the OLS line
- Must be context-specific, specify units, be clear on causality, and assess magnitude (big or small)
- For example, in drug trial: $\hat{y}_i = 3.44 + 0.48x_i$
  - How to *interpret* 0.48?
    - *We estimate that an extra mg of drug causes [yields/results in] about an extra half hour (29 mins) of sleep on average.*
  - How to *interpret* 3.44?
    - *We don't: 0 mg is outside the range of the data and we should not extrapolate in this case because 0 mg is distinct.*

Causality and Observational Data?

“Glassdoor [used] records of 293 companies across 13 industries between 2008 and 2018. It studied the link between employee satisfaction, based on its own ratings, and the American Customer Satisfaction Index, a benchmark gauge of shoppers’ sentiment.”



Source: Glassdoor †Glassdoor rating, maximum=5

*American Customer Satisfaction Index, maximum=100

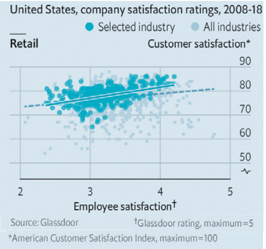
As an interpretation of the dashed OLS line, what’s wrong with this?

“A one-point improvement in Glassdoor’s rating (on a five-point scale) translated into a statistically significant 1.3-point increase in customer satisfaction (rated from zero to 100).” What does it get right?

The Economist, 2019, “How to keep your customers happy: One way is to ensure your workers are, too” https://www.economist.com/business/2019/08/22/how-to-keep-your-customers-happy

Descriptive Interpretations

OLS line: $\hat{y} = 62.83 + 4.73x$ (solid line) Interpretation in 3 sentences?



Source: Glassdoor †Glassdoor rating, maximum=5

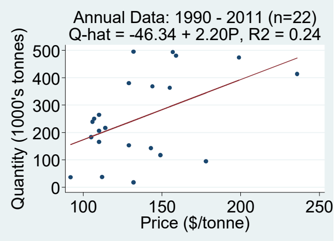
*American Customer Satisfaction Index, maximum=100

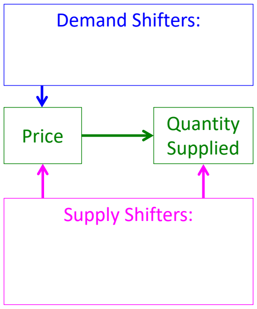
For U.S. retail firms from 2008 to 2018 tracked by Glassdoor, those with employee satisfaction that is 1 point higher on a five-point scale have customer satisfaction that is 4.7 points higher on a 100 point index on average.

This is a modest difference as having 1 point higher employee satisfaction is huge (roughly moving from 10th to 90th percentile).

We cannot infer causality. We simply describe the pattern.

Prices & Production: Manitoba Corn Farms





Demand Shifters:

Price

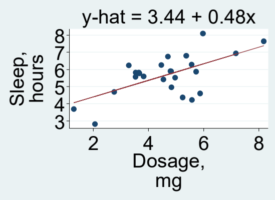
Quantity Supplied

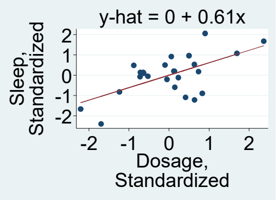
Supply Shifters:

Which kind of data are these?

Source: Manitoba gov’t: http://www.gov.mb.ca/agriculture/statistics/pdf/crop_grain_corn_sector.pdf

Interpretation of Correlation, Revisited





$$\boxed{b = r\frac{s_y}{s_x}}$$

What (else) is 0.61?

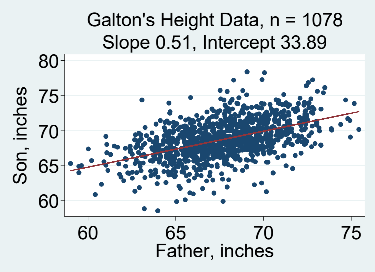
See pp. 177 – 178 of the textbook.

Least Squares (OLS) = Regression Analysis

Sir Francis Galton 1822-1911

- Sweet peas: Compare size of initial seed to seeds produced by plant once mature
- Intelligence: Compare children to their parents
- Heights: Compare children to their parents

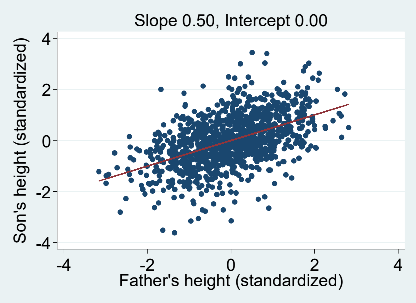
Regression Towards the Mean



For sons: mean is 68.7 inches and s.d. is 2.8 inches.

For fathers: mean is 67.7 inches and s.d. is 2.7 inches.

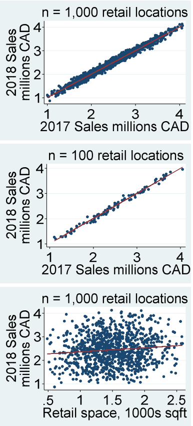
Suppose Standardized Heights



Because x and y are each standardized, slope (b) equals coefficient of correlation (r) and $-1 \leq r \leq 1$. Hence, a 1 s.d. increase in x cannot be associated with more than 1 s.d. change in y.

Variance and Scatter: Not about n

- Recall $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$
  - Both numerator (SST) & denominator rise with n: there is no net effect
  - How much y varies in the population is the key
- Scatter is about our trouble predicting y: it depends on how useful x is, not on n



Residuals (error): $e_i = y_i - \hat{y}_i$

- Constant term → mean of residuals is 0
- Estimated s.d. of residuals; Root MSE (Mean Square Error):

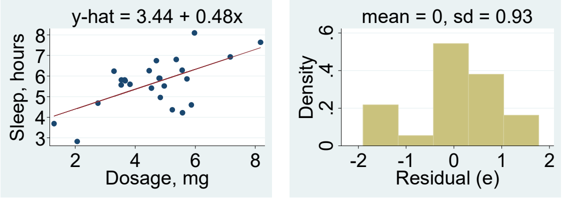
$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(e_i - 0)^2}{n-2}}$$

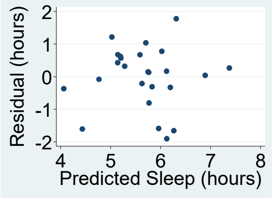
"standard error of estimate" $s_e = \sqrt{\frac{SSE}{n-2}}$

$$\hat{y}_i = 3.44 + 0.48x_i$$

i	x_i (mg)	y_i (hrs)	$\hat{y}_i$ (hrs)	e_i (hrs)
1	5.9	4.6	6.3	-1.7
2	3.5	5.8	5.1	0.7
3	7.2	6.9	6.9	0.0
4	3.6	5.8	5.2	0.6
…	…	…		
25	8.2	7.6	7.4	0.2

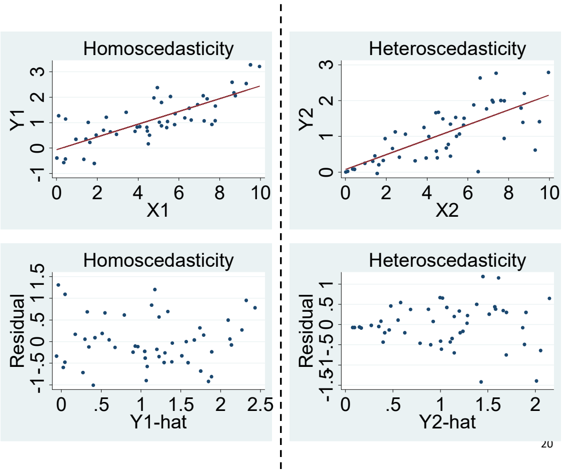
Used up 2 degrees of freedom: $e_i = y_i - a - bx_i$





Homoscedasticity: The variance of the residual is constant

Homoscedasticity means that it makes sense to talk about *the* standard deviation of the residual

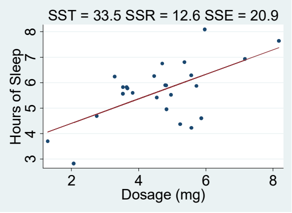


Analysis of Variance (ANOVA): Fit

- ANOVA: How total variability of y relates to x versus everything else
- Total sum of squares: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Regression sum of squares: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Sum of squared errors: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- $SST = SSR + SSE$

Units of measurement?

What happens as n increases?

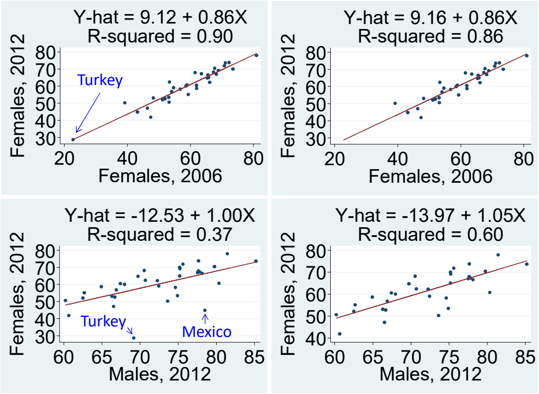


If x had nothing to do with y, what would the SSR be? SSE?

If x explained y perfectly, what would the SSE be? SSR?

R^2 (Coefficient of Determination): A Measure of Fit

- $R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$
  - Fraction of total variation in dependent variable (y) explained by indep. variable (x)
    - $1 - R^2$: unexplained variation in y
  - For simple regression: $R^2 = (r)^2$
    - For example, Galton's data: $R^2 = \frac{2144.6\ inches^2}{8532.6\ inches^2} = 0.25$ Twenty five percent of the variation in height across sons is explained by variation in the heights of their fathers

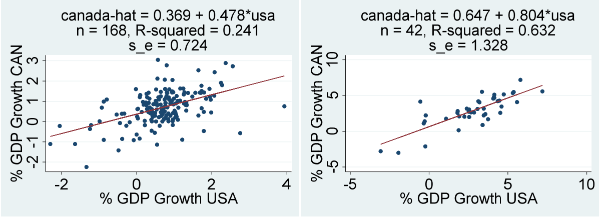


OECD data, again (n = 34 countries)

Summary Values

- If the available data contain summaries of a larger data set (e.g. means, totals, etc.) this affects the regression analysis
- Example: Suppose an analyst knows that the Canadian and U.S. economies are closely linked and regresses Canadian GDP growth (%) on U.S. GDP growth
  - Does it matter if it is annual or quarterly data?

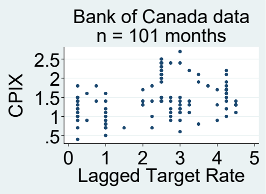
1971 – 2012 Data (OECD)



Which is the quarterly data? Annual data?

Very Brief Review: Lectures 2 & 4

- $\bar{y} = 1.433$ $\boxed{\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}}$
- $s_y = 0.497$ $\boxed{s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$
- $\bar{x} = 2.453$
- $s_x = 1.395$
- $s_{xy} = 0.215$ $\boxed{s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}$



- $r = 0.309$ $\boxed{r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n-1}}$

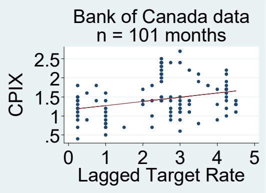
Very Brief Summary: Lecture 5

- $\hat{y}_i = 1.16 + 0.11x_i$

$\boxed{b = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}}$ $\boxed{a = \bar{y} - b\bar{x}}$

- Residuals: $\boxed{e_i = y_i - \hat{y}_i}$
  - E.g. Jan. 2013: $x_{101} = 1$, $y_{101} = 0.5$, $\hat{y}_{101} = 1.27$, $e_{101} = -0.77$ $(= 0.5 - 1.27)$
- s.d. of resids: $s_e = 0.4748$

$$\boxed{s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}}$$



- $SST = 24.7 = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $SSR = 2.4 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- $SSE = 22.3 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- $R^2 = 0.10 = SSR/SST$

If standardized both variables, slope of regression line?

Top 10 reasons why *interpreting* an OLS line is NOT like a geometric line

1. Most/all data do NOT lie on an OLS line: say "on average" to recognize scatter. A geometric line maps from x to y exactly.
2. Even for a sample size of two, where the OLS and geometric lines would overlap, when interpreting the OLS line you must remember the huge sampling error for such a small sample.
3. Consider how well an OLS line fits data with statistics like R^2.
4. The OLS slope is the coefficient of correlation if the data are standardized. This makes no sense for a geometric line.
5. A geometric line implies causality but an OLS line can only be interpreted causally with experimental data; for observational data, slope likely suffers an endogeneity bias.

Top 10 reasons why *interpreting* an OLS line is NOT like a geometric line

6. For a geometric line, can write y in terms of x or x in terms of y: $y = a + bx$ equivalent to $x = -\frac{a}{b} + \frac{1}{b}y$. For an OLS line if you switch x and y variables you get a *different* line, not just a re-written one.
7. While an intercept has a simple interpretation for a geometric line (value of y when x is zero), for an OLS line the intercept often has no meaning or should not be interpreted because zero is well beyond the range of the data.
8. An OLS line is not robust to outliers: report results with and without the outlier(s). In contrast, there is only one geometric line and all points lie on it.

Top 10 reasons why *interpreting* an OLS line is NOT like a geometric line

9. An OLS line may describe variables where one or both have been non-linearly transformed and the "slope" is *not* interpreted as a slope. This is the point of the required reading "Logarithms in Regression Analysis with Asiaphoria."
10. OLS lines require substantial expertise to interpret properly.

But, this doesn't mean nothing translates. Changes in units of measurement change a line in perfectly predictable ways:

*If y-hat = 107.07 – 4.20*x when x measured in hours,*
*then y-hat = 107.07 – 0.07*x when x is in minutes*

*If y-hat = 20.48 + 0.61*x when y measured in \$100's,*
*then y-hat = 2.048 + 0.061*x when y is in \$1,000's*

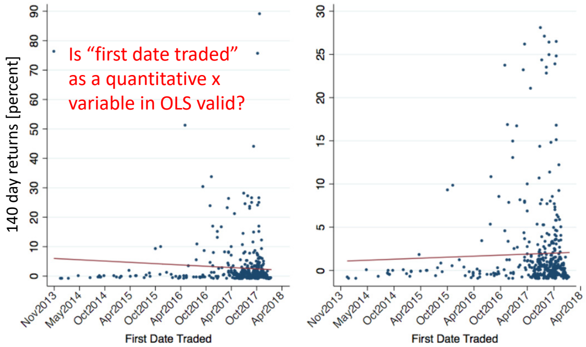
Review all parts of question 10 on the quiz in the *Quiz and Prerequisite Review* (p. 5).

“Initial Coin Offerings: Financing Growth with Cryptocurrency Token Sales”

ABSTRACT: Initial coin offerings (ICOs) are sales of blockchain-based digital tokens associated with specific platforms or assets. Since 2014 ICOs have emerged as a new financing instrument, with some parallels to IPOs, venture capital, and pre-sale crowdfunding. We examine the relationship between issuer characteristics and measures of success, with a focus on liquidity, using 453 ICOs that collectively raise $5.7 billion. We also employ propriety transaction data in a case study of Filecoin, one of the most successful ICOs. We find that liquidity and trading volume are higher when issuers offer voluntary disclosure, credibly commit to the project, and signal quality.

Howell et. al. (2018) http://www.nber.org/papers/w24774

Figure 9: Cumulative Returns at 5 Months
[n = 444 ICOs, Fitted line in red]



Is “first date traded” as a quantitative x variable in OLS valid?

Note: Cumulative returns between the start of trading and 140 days (5 months) subsequently. Right panel excludes observations w/ returns > 30 [outliers].