Univariate Regression

(Part 2 – Estimation and Measures of Fit)

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w - Spring 2022

Brooklyn College
Feb 28, 2022

Outline

1. **Probability framework for linear regression**
2. **The ordinary least squares (OLS) estimator and the sample regression line**
3. **Measures of fit of the sample regression**
4. The least squares model assumptions
5. The sampling distribution of the OLS estimator

Linear regression lets us estimate the population regression line and its slope.

- The population regression line is the expected value of *Y* given *X*.
- The slope is the difference in the expected values of *Y*, for two values of *X* that differ by one unit
- The estimated regression can be used either for:
  - causal inference (learning about the causal effect on *Y* of a change in *X*)
  - prediction (predicting the value of *Y* given *X*, for an observation not in the data set)

Probability framework for linear regression

- *Population*

  population of interest (ex: all possible school districts)

- *Random variables*: Y, X

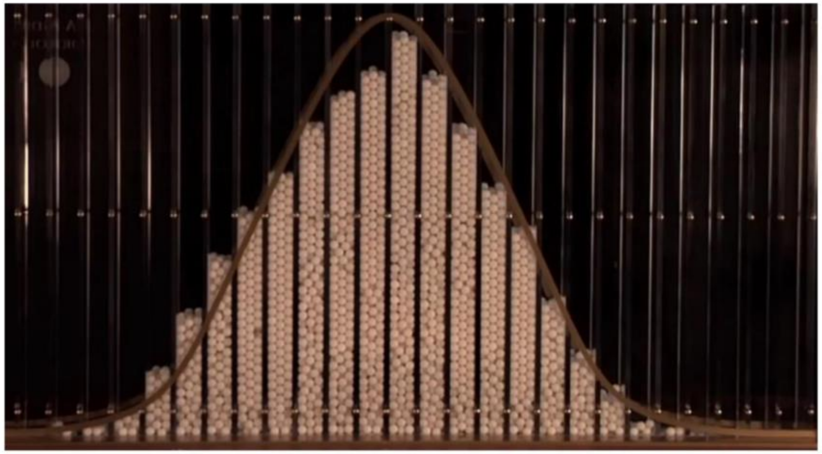
  Ex: (*Test Score*, *STR*)

- *Joint distribution of* (Y, X)

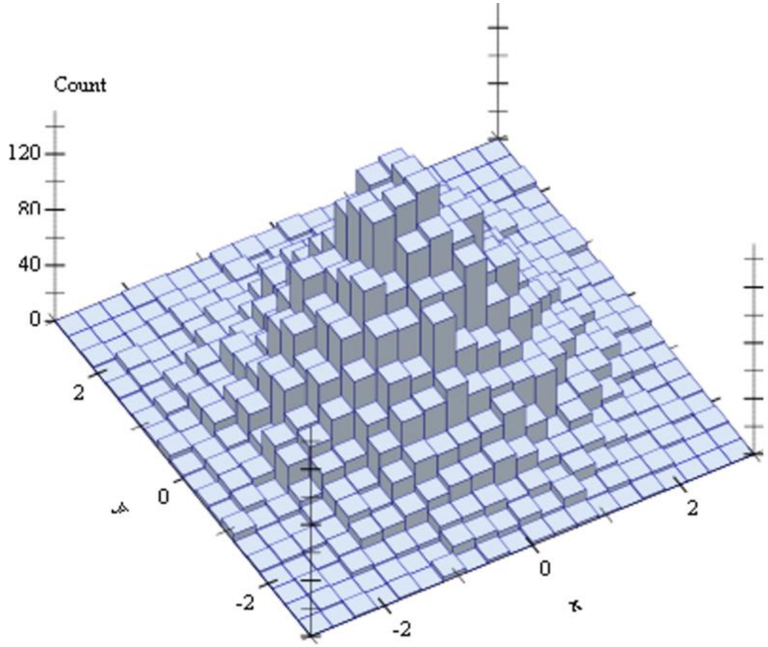
  The key feature is that we suppose there is a linear relation in the population that relates X and Y; this linear relation is the "population linear regression"

Joint distribution visualization

Histogram (univariate)

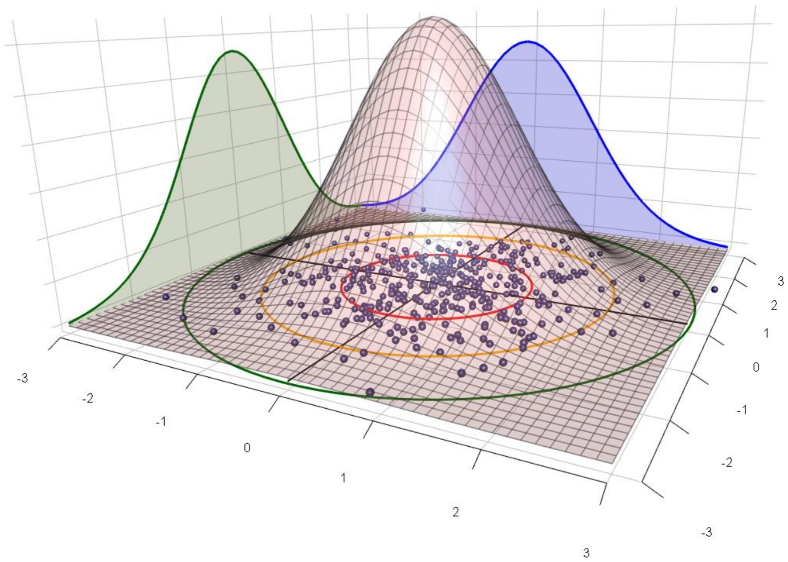


Histogram (bivariate)

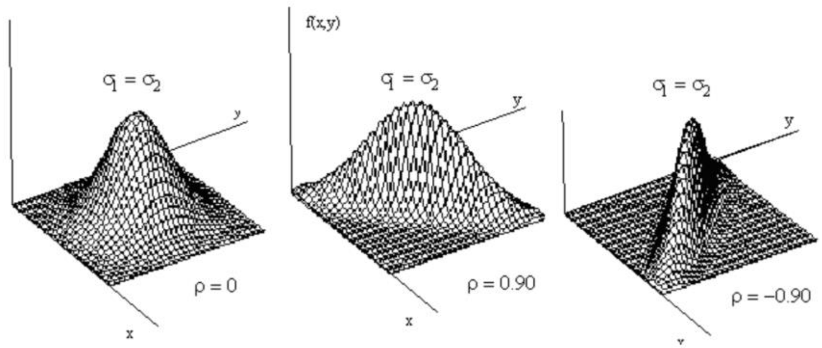


Joint distribution visualization (cont'd)

Sampling from bivariate distribution (correlation =0)



Bivariate distributions (0, positive, and negative correlation)



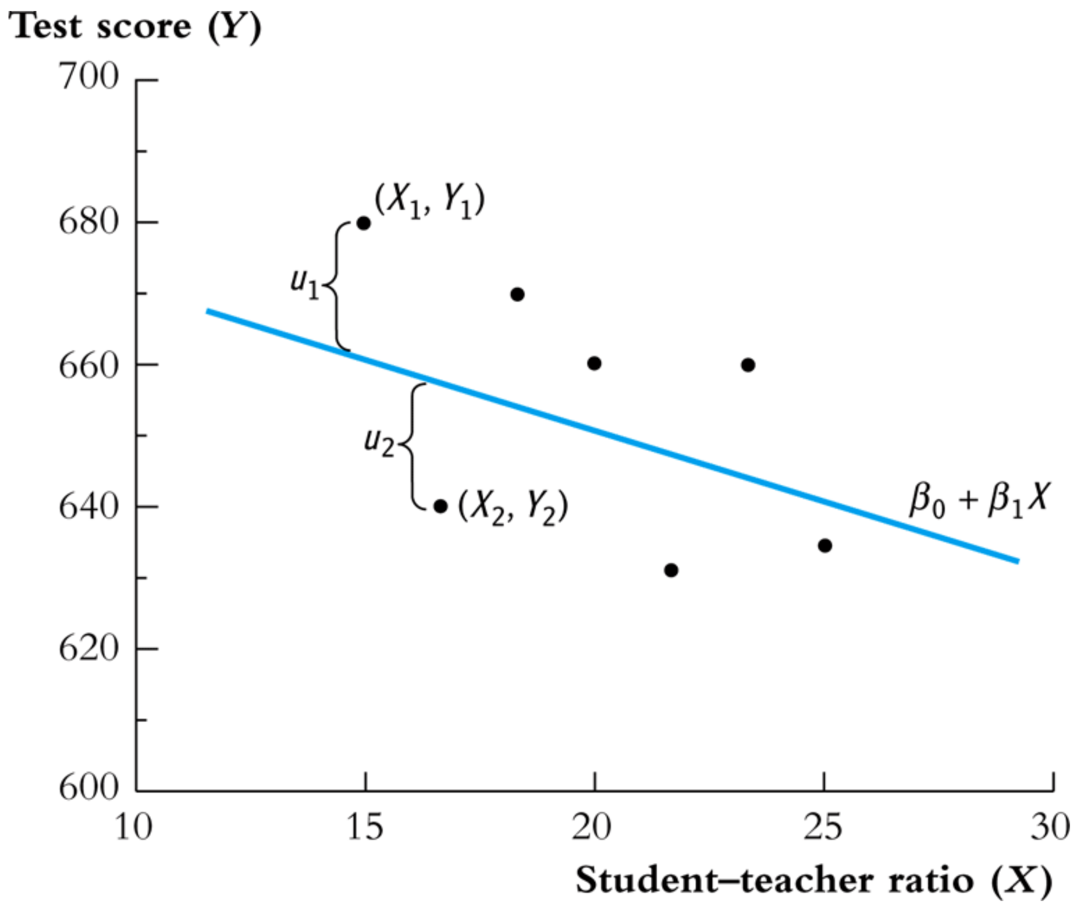
https://demonstrations.wolfram.com/TheBivariateNormalAndConditionalDistributions/

The Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots, n$$

- We have n observations, (X_i, Y_i), $i = 1,.., n$.
- X is the ***independent variable*** or ***regressor***
- Y is the ***dependent variable***
- β_0 = ***intercept***
- β_1 = ***slope***
- u_i = "***error term***"
- The error term consists of omitted factors, or possibly measurement error in the measurement of Y. In general, these omitted factors are other factors that influence Y, other than the variable X

The population regression model in a picture: *Observations* on *Y and X (n = 7); the population regression line; and the* regression error (the "error term"):



What are some of the omitted factors in this example?
e.g. parental involvement, outside learning opportunities (extra math class,..), home environment conducive to reading, etc.

The Ordinary Least Squares Estimator

We will focus on the least squares ("*ordinary least squares*" or "*OLS*") estimator of the unknown parameters β_0 **and** β_1.

The OLS estimator solves,

$$\min_{b_0,b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem can be solved using calculus (App. 4.2).
- **The result is the OLS estimators of** $\boldsymbol{\beta_0}$ **and** $\boldsymbol{\beta_1}$ **(**$\boldsymbol{\hat{\beta}_0}$ **,** $\boldsymbol{\hat{\beta}_1}$ **or** $\boldsymbol{b_0}$ **,** $\boldsymbol{b_1}$**)**

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.8}$$

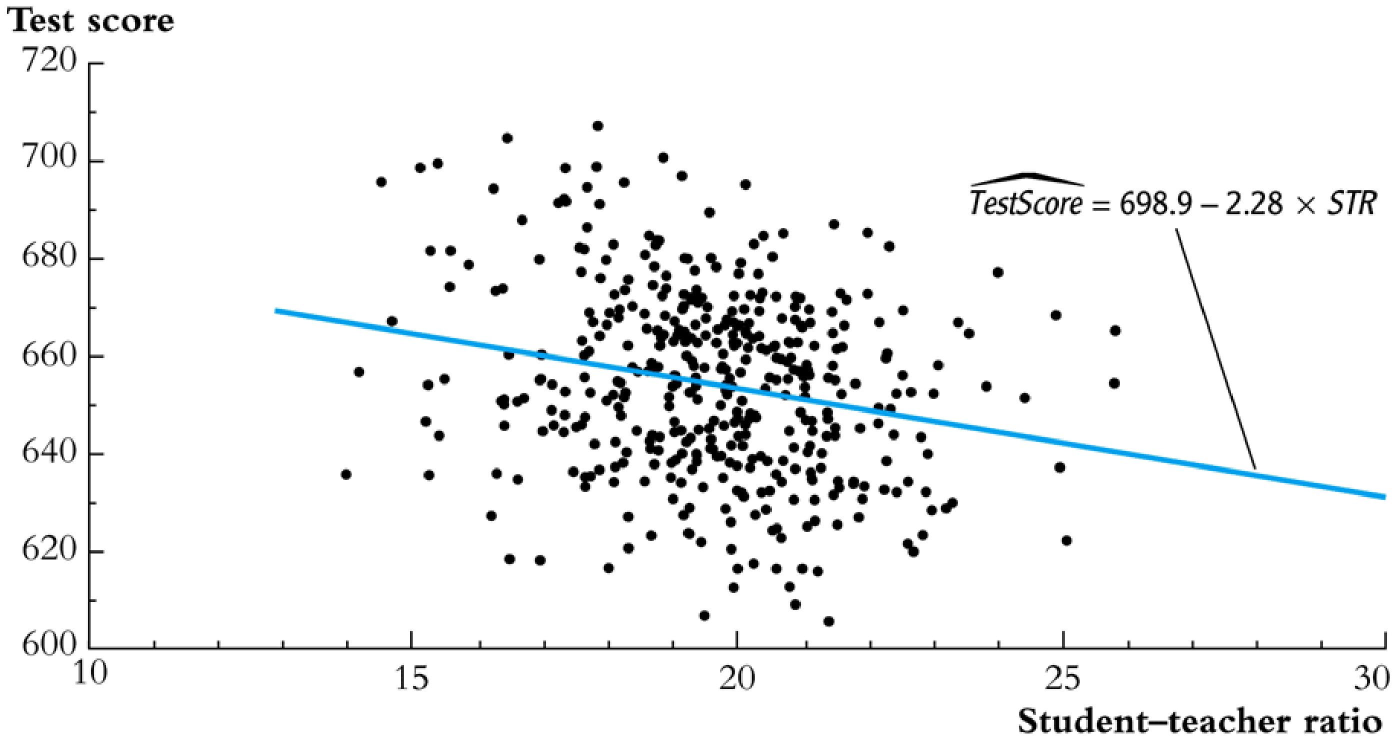
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

Application to the California *Test Score vs Class Size* data



- Estimated slope $= \hat{\beta}_1 = -2.28$
- Estimated intercept $= \hat{\beta}_0 = 698.9$
- Estimated regression line: $\widehat{TestScore} \quad = 698.9 - 2.28 \times STR$

Interpretation of the estimated slope and intercept

- $\widehat{TestScore} = 698.9 - 2.28 \times STR$

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\dfrac{\Delta E(Test\ score|STR)}{\Delta STR} = -2.28$

- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. But this interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

Example

For the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ find the OLS estimators $\hat{\beta}_0$, $\hat{\beta}_1$ given the following data sample:

Y_i	X_i
4	1
5	4
7	5
12	6

Measures of Fit (SW Section 4.3)

Two regression statistics provide complementary measures of how well the regression line "fits" or explains the data:

- The ***regression R^2*** (aka "coefficient of determination) measures the fraction of the variance of *Y* that is explained by *X*; it is unitless and ranges between zero (no fit) and one (perfect fit)
- The ***standard error of the regression* (*SER*)** measures the magnitude of a typical regression residual in the units of *Y*.

The *regression* R^2 is the fraction of the sample variance of Y_i "explained" by the regression.

$Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual
$\rightarrow$ sample var (Y) = sample var$(\hat{Y}_i)$ + sample var$(\hat{u}_i)$
$\rightarrow$ total sum of squares = "explained" SS + "residual" SS

Definition of R^2: $$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \leq R^2 \leq 1$
- For regression with a single X, R^2 = the square of the correlation coefficient between X and Y

The Standard Error of the Regression (SER)

The *SER* measures the spread of the distribution of u. The *SER* is (almost) the sample standard deviation of the OLS residuals:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

The second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$.

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

The *SER*:

has the units of u, which are the units of Y

measures the average “size” of the OLS residual (the average “mistake” made by the OLS regression line)

The ***root mean squared error*** (*RMSE*) is closely related to the *SER*:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

This measures the same thing as the *SER* – the minor difference is division by $1/n$ instead of $1/(n – 2)$.

Example (continued)

For the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, find the coefficient of determination (R^2) given the following data sample:

Y_i	X_i
4	1
5	4
7	5
12	6