Sabancı Üniversitesi FACULTY OF ARTS AND SOCIAL SCIENCES

Introduction to OLS and Regression Analysis

Berfin Baydar

SICSS-Istanbul 2023

Faculty of Arts and Sciences, Sabancı University

July 4, 2023

Table of Contents

The Simple Regression Model

- Regression: Statistical models to explore the relationship between a response variable and some explanatory variables.
- A regression analysis is about the expected value of Y conditional on a single X (in the bivariate case) or multiple (in the multivariate case) Xs.
  - Explaining Y in terms of X
  - Studying how Y varies with changes in X
- Two variable / Bivariate Linear / Simple Regression Model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- A note on terminology:
  - We run the regression of y on x
  - We regress y on x
  - We always regress the dependent variable on the independent variable!

The Simple Regression Model Cont'd.

$$y = \beta_0 + \beta_1 x + \epsilon$$

- y: dependent variable, explained variable, response variable, predicted variable, regressand
- x: independent variable, explanatory variable, control variable, predictor variable, regressor, covariate
- ϵ: error term, disturbance, stochastic/random component (unobserved factors other than x that affect y)
- β_0: intercept parameter, constant (average value of y when x is equal to zero)
- β_1: slope parameter, slope coefficient for X_1

Linear Relationship between x and y

- Given values of explanatory variables, you may predict the values of the dependent variable.
- If $\Delta\epsilon = 0$ then $\Delta y = \beta_1 \Delta x$
- This means that when holding other things fixed (ceteris paribus), x has a linear effect on y:
  - The linearity between x and y means a one-unit change in the x has the same effect on y regardless of the initial value of x.
  - A one-unit increase in x changes the expected value of y by the amount of β_1 (the slope).

Estimating the β_0 and β_1

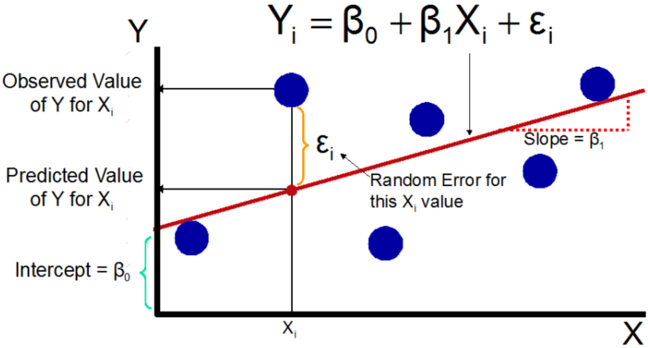
- β_0 and β_1 are estimated using the data.
- The estimates of parameters are shown by “hats”
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ –> estimates of β_0 and β_1
- Once we obtain the estimated values of coefficients, we can draw our regression line!

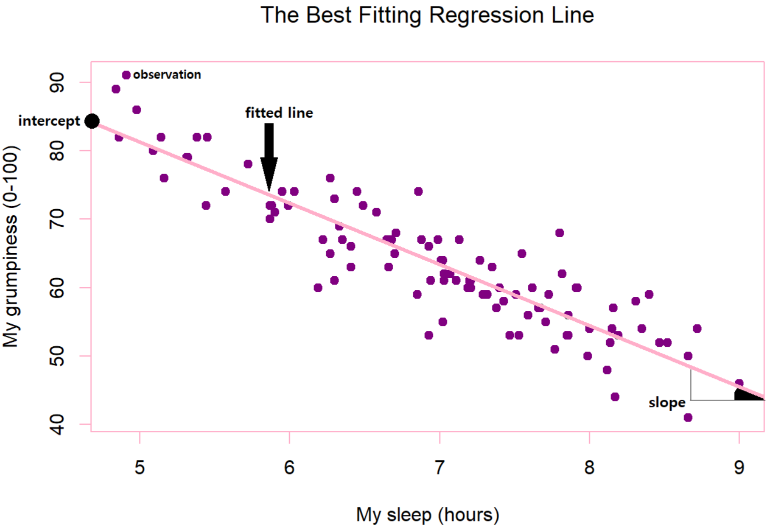
Ordinary Least Squares Estimator

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{y}_i$ is the value we predict for y when $x = x_i$ for the given intercept and slope.
- There is a fitted value for each observation in the sample. Each fitted value of $\hat{y}_i$ is on the OLS regression line.
- We use “least squares” to estimate the line’s intercept and slope parameters.
- A line provides the best fit to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.
  - So, it is called “least” because the OLS estimates minimize the sum of squared residuals (SSR).

Explaining Coefficients

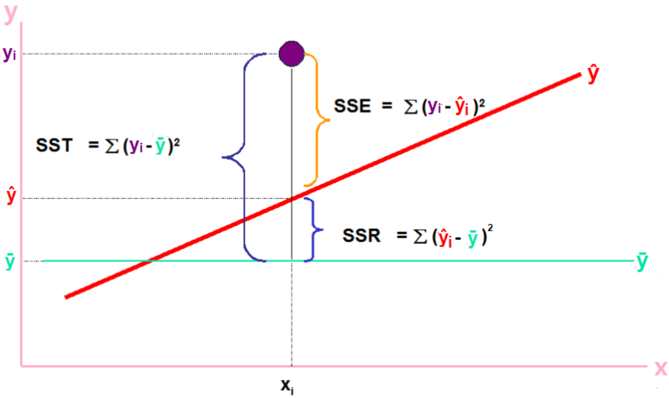

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X


The Best Fitting Regression Line
observation
intercept
fitted line
slope
90
80
70
60
50
40
My grumpiness (0-100)
5
6
7
8
9
My sleep (hours)

Explaining Squares

- **Sum of Squares Total:** is the sum of squared differences between the observed dependent variables and the overall mean. Think of it as the dispersion of the observed variables around the mean—similar to the variance in descriptive statistics. But SST measures the total variability of a dataset.
  - $SST = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$
- **Sum of Squares Error:** is the difference between the observed and predicted values. Regression analysis aims to minimize the SSE—the smaller the error, the better the regression's estimation power.
  - $SSE = \sum_{i=1}^{N}(Y - \hat{Y})^2$
- **Sum of Squares Regression:** is the sum of the differences between the predicted value and the mean of the dependent variable. In other words, it describes how well our line fits the data.
  - $SSR = \sum_{i=1}^{N}(\hat{Y} - \bar{Y})^2$
- The total variability of the dataset is equal to the variability explained by the regression line plus the unexplained variability, known as error.
  - $SST = SSE + SSR$

Explaining Squares


y
y_i
$\hat{y}$
$\bar{y}$
$\hat{y}$
SSE $= \Sigma (y_i - \hat{y}_i)^2$
SST $= \Sigma (y_i - \bar{y})^2$
SSR $= \Sigma (\hat{y}_i - \bar{y})^2$
$\bar{y}$
x
x_i

Deriving the OLS Estimator

- Assuming that the mean of ϵ is zero and that X_1 and ϵ are not correlated:
  - $E[\epsilon] = 0$
  - $E[X_1|\epsilon] = 0$
    - $Y = \beta_0 + \beta_1 X_1 + \epsilon$
- OLS residuals always sum to zero!
- The population function of Y is assumed to be $Y = \beta_0 + \beta_1 X_1 + \epsilon$, and the residual for the ith observation is $\epsilon_i = Y_i - \beta_0 - \beta_1 X_{1i}$
- We can thus restate our first two assumptions in terms of Y
  - $E[Y_i - (\beta_0 + \beta_1 X_1 i)] = 0$
  - $E[X_{1i}|(Y_i - (\beta_0 + \beta_1 X_{1i}))] = 0$
  - where the first condition is that the expected mean of the disturbances and the second is that the covariance of X_1 (or all Xs in the multivariate case) and the disturbances are both equal to zero.

Deriving the OLS Estimator Cont'd.

- For a sample of size N, the estimates of the population parameters $\beta_0(\hat{\beta}_0)$ and $\beta_1(\hat{\beta}_1)$ can be stated as:
    - $\frac{\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_i X_i)}{N} = 0$
    - $\frac{\sum_{i=1}^{N} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_i X_i)}{N} = 0$
    - Because $(k+1)x1$ vector of ordinary least squares estimates, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$, minimizes the sum of squared residuals over all possible $(k+1)x1$ vectors; and for $\hat{\beta}$ to minimize the SSR, it must solve the first order condition.
    - The first order conditions above are equivalent to taking the partial derivatives of the sum of squares with respect to the unknowns, which is yet another way of deriving the OLS estimator.

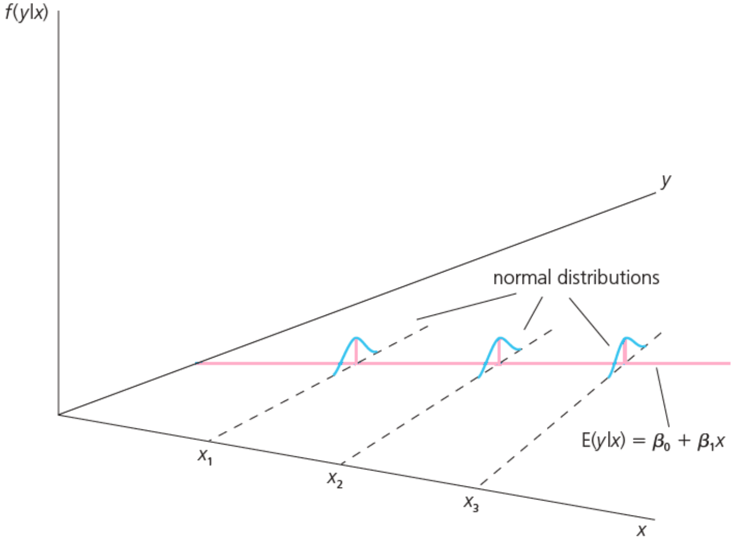
Deriving the OLS Estimator Cont'd.

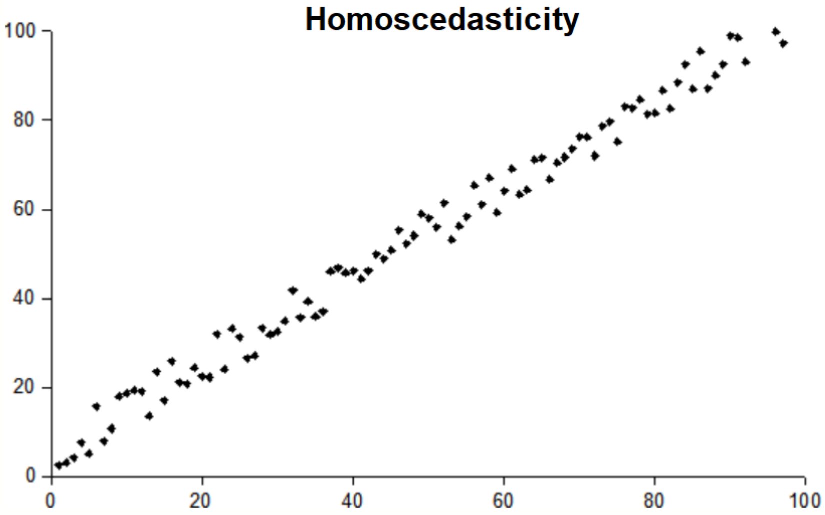
- If $X'\hat{\epsilon} = 0$ holds, then the following properties exist:
  - Each X_k is uncorrelated with ϵ
  - If the sum of residuals is zero, then the mean of the residuals must be equal to zero (the sum divided by N).
  - The regression line passes through the means of observed Y and X (or $\bar{X}\hat{\beta}$).
    - $\bar{Y} - \bar{X}\hat{\beta} = 0$ implies the intersection of the means with the regression line.
    - This is simply because we solve the equation by minimizing the sum of squared residuals.

Properties of OLS

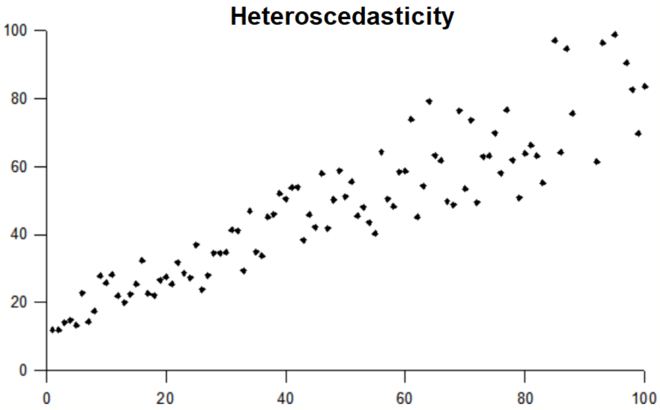
- **Linear in parameters:**
  - $Y = \beta_0 + \beta_1 x + \epsilon$
- **No perfect collinearity/Sample Variation in the Explanatory Variable:**
  - None of the independent variables is constant, and there are no exact linear relationships among the independent variables.
- **Zero conditional mean:**
  - The error ϵ has an expected value of zero given any value of the explanatory variable.
    $$E(\epsilon|x) = 0$$
- **Homoscedasticity and No Serial Correlation:**
  - Homoscedasticity means that the error term has a constant variance across all values of the independent variables.
    $$Var(\epsilon|x) = \sigma^2$$
  - The error terms should not exhibit serial correlation, meaning they should be independent of each other.
- **Normality of Errors:**
  - Residuals are independent and identically distributed. The error term is assumed to follow a normal distribution.

The Homoscedastic Normal Distribution




Homoscedasticity
100
80
60
40
20
0
0
20
40
60
80
100

- If the error variance depends on the predictor, we say that the error is heteroskedastic.
- In this case, we use heteroskedasticity-robust standard errors.



Model Accuracy

- After obtaining the estimates of the population parameters, it is important to assess how well the model fits the data. This evaluation process is commonly known as the **goodness-of-fit test**.
- Three commonly used measures to assess the quality of a linear regression fit are:
  - **Residual Standard Error (RSE):** Estimates the standard deviation of the residuals, indicating the average distance between observed and predicted values. By calculating the ratio of RSE to the average value of the outcome variable, you can obtain a relative measure of prediction error. Lower values indicate a better fit.
  - **R-squared (R^2):** Represents the proportion of variance in the dependent variable explained by the independent variables. It ranges from 0 to 1, where higher values indicate a better fit, but it should be interpreted alongside other factors.
    - As the value of R^2 tends to increase when more predictors are added in the model, you should mainly consider the adjusted R-squared, which is a penalized R^2 for a higher number of predictors.
  - **F-statistic:** Tests the overall significance of the regression model by comparing the variability explained by the model to the unexplained variability. A significant F-statistic suggests a better fit.

Ommitted Variable Bias

- The problem of excluding a relevant variable or underspecifying the model
  - This causes the OLS estimators to be biased.
  - Remember, correlation between a single explanatory variable and the error generally results in all OLS estimators being biased!
- The bias in β can occur in two ways:
  - Omitted variable bias due to correlation: If the omitted variable is correlated with both the dependent variable and the included independent variable(s), the estimated coefficient for the included variable(s) will be biased. The estimated coefficient will incorporate the combined effects of the included variable(s) and the omitted variable, thus leading to an incorrect interpretation of the relationship between the included variable(s) and the dependent variable.
  - Omitted variable bias due to confounding: If the omitted variable is a confounding variable, meaning it affects both the dependent variable and the included independent variable(s), the estimated coefficient for the included variable(s) will be biased. The confounding variable introduces a spurious relationship between the included variable(s) and the dependent variable, which results in biased estimates.

Violation of Exogeneity Assumption

- The exogeneity assumption implies that the mean of the error term does not depend on the predictors or explanatory variables included in the model.
  - The unobserved determinants of the outcome variable, which are contained in the error term, should be uncorrelated with all the observed predictors.
- In observational studies, the exogeneity assumption may be violated.
- How can we address this problem of unobserved confounding in observational studies?
  - Compare the treated units with similar control units.
- The assumption of no unobserved confounding has several different names: unconfoundedness, selection on observables, and no omitted variables.
- In the linear regression model framework, we can solve this problem by measuring these confounders and including them as additional predictors in the model.

Panel Data and Fixed Effect Regression

- Panel data, also known as longitudinal or repeated measures data, consists of observations on multiple entities (such as individuals, firms, or countries) over time.
    - In panel data analysis, fixed effects refer to a type of estimation technique that accounts for individual-specific or time-invariant heterogeneity in the data.
    - These individual-specific effects, also known as fixed effects or entity-specific effects, can capture unobserved factors that are unique to each entity.
- Fixed-effects allow us to estimate the relationship between the independent variables and the dependent variable while controlling for individual-specific effects.
    - It involves including individual-specific dummy variables (also called indicator variables or entity dummies) in the regression model.
    - These dummy variables take a value of 1 if the observation belongs to a specific entity and 0 otherwise.
    - The coefficient associated with each dummy variable captures the average effect of that specific entity on the dependent variable, after controlling for other independent variables.

Reporting Regression Results

- The estimated OLS coefficients should always be reported.
  - You should interpret the estimated coefficients (which often requires knowing the units of measurement of the variables).
- The standard errors should always be included along with the estimated coefficients.
  - Having standard errors makes it easier to compute confidence intervals.
- The R-squared from the regression should always be included.
  - In addition to providing a goodness-of-fit measure, it makes calculation of F-statistics for exclusion restrictions simple.
- The number of observations used in estimating any equation should be included.

Thank you for listening!