# Chapter 10: Regression Analysis: Estimating Relationships

## Introduction

Regression is the study of relationships between variables. There are 2 potential objectives: understanding how the world operates + making predictions.  
There are 2 types of data: cross-sectional + time series  
Time series variables are potentially related to their own past values: autocorrelation.  
Simple regression: single explanatory variable  
Multiple regression: several explanatory variables

2 types of relationships being studied: linear vs. non-linear relationships  
Linear **regression** studies straight-line **relationships** + linearizes nonlinear **relationships** by using suitable mathematical transformations

## Scatterplots: graphic relationships

Graphical plot of 2 variables.

Correlation ≠ causation: you can’t always be sure which direction the causation goes + you can almost never rule out the possibility that some other variable is causing the variation.

### 2.a. Linear vs. Nonlinear Relationships

Scatterplot => nonlinear curvature: linear regression might still be applicable after an appropriate transformation of the date.

### 2.b. Outliers

Observation that falls outside of the general pattern of the rest of the observations.  
- If an outlier is clearly not member of the population of interest -> delete  
- If not clear -> run regression analysis with and without them and compare the results.

### 2.c. Unequal Variance

Variance of the dependent variable depends on the value of the explanatory variable. The scatterplot has a distinct fan shape.

### 2.d. No relationship

The scatterplot appears as a shapeless swarm of points.

## Correlations: Indicators of Linear Relationships

Correlations are numerical summary measures that indicate the strength of linear relationships between pairs of variables.  
Correlation: covariance / product of sample standard deviations  
The magnitude of the covariance depends on the units of measurement.  
The magnitude of correlation is always between -1 and 1.  
No linear relationship: 0

## Simple Linear Regression

### 4.a. Least Squares Estimation

The line that makes the vertical distances from the points to the line as small as possible.  
Fitted value: predicted value of the dependent variable.  
Residual value: difference between fitted and observed.  
Under the line: negative residual  
Above the line: positive residuals  
Fundamental equation for Regression: Observed = Fitted + Residual  
Least squares line: line that minimizes the sum of the squared residuals.

### 4.b. Standard Error of Estimate

Standard error of estimate: se

2/3 of the fitted ^y values are within 1 standard error of the mean.  
95% of the fitted ^y values are within 2 standard errors of the mean.  
The standard error of estimate indicates the level of accuracy of predictions made from the regression equation. The smaller it is, the more accurate predictions tend to be.

### 4.c. The percentage of variation explained: R-square

R²: fraction of the variation explained by the regression line.  
The only way to increase R² is to use better and/or more explanatory variables.  
Square of the correlations between the observed Y-values and the fitted ^y-values.

## Multiple Regression

2 possible approaches to obtaining improved fits: Examine a scatterplot of residuals for nonlinear patterns and make appropriate modifications to the regression equation.

### 5.a. Interpretation of regression coefficients

The intercept a is the expected value of Y when all of the X’s equal 0.  
Each slope coefficient b is the expected change in Y when the corresponding X changes by one unit.

### 5.b. Interpretation of Standard Error of Estimate & R²

Standard error of estimates: standard deviation of residuals.  
Adjusted R²: to avoid adding variables just to inflate R²; no direct interpretation.

## Modelling possibilities

### 6.a. Dummy variables

Categorical variables that cannot be measured on a quantitative scale = Indicator or 0-1 variables.  
When a categorical variable has only two categorie: ex. Male and female.  
When a categorical variable has more than two categories: make a dummy variable for each category and use 1 as yes and 0 as no.  
Always use one fewer dummy than the number of categories for any categorical value and use the last one as a reference value.

### 6.b. Interaction variables

An interaction variable is the product of two explanatory variables. You can include such a variable in a regression equation if you believe the effect of one explanatory variable on Y depends on the value of another explanatory variable.

### 6.c. Nonlinear transformations

Curvature detected in scatterplots or economic considerations.

* Natural logarithm
* Square root
* Reciprocal
* Square

Straighten out the points in a scatterplot.  
Any coefficient b can now be interpreted as the approximate percentage change in Y when the corresponding X increases by one unit.  
- R² values of Y and Log(Y) are **not** comparable.  
- Se values of Y and Log (Y) are **not** comparable.  
- To interpret any term in the form of bX in log equation, express b as a percentage.

### 6.d. Constant Elasticity Relationships

Firm grounding in economic theory. Multiplicative relationships. The dependent variable is expressed as a product of explanatory variables raised to powers.