Linear Regression

Ani Katchova

Linear Regression Overview

- Linear regression examples
- Linear regression model
- Estimated regression line
- Single versus multiple regression
- Coefficients and marginal effects
- Goodness of fit (R-squared)
- Hypothesis testing for coefficient significance
    - t-test for a single coefficient significance
    - F-test for multiple coefficients significance

Linear Regression

Linear regression examples

- Explain student grades using the number of hours studied
- Explain the effect of education on income
- Explain the effect of the number of bedrooms on house prices
- Explain the effect of the recession on stock prices

Linear regression set up

- Regression analysis does not establish a cause-and-effect relationship, just that there is a relationship.
- The cause-and-effect relationship must be determined using a theoretical model or a logical reason.
- The dependent variable is a continuous variable.
- The independent variables can take any form - continuous or discrete or indicator variables.
- The simple linear regression model has one independent variable.
- The multiple linear regression model has two or more independent variables.

Linear regression model

- The linear regression model describes how the dependent variable is related to the independent variable(s) and the error term:

$$y = \beta_0 + \beta_1 x_1 + u$$

or

$$y = x'\beta + u$$

- y is the dependent variable (explained, predicted, or response variable)
- x is the independent variables (control variables or regressors)
- β are unknown parameters to be estimated
  - β_0 is the intercept
  - β_1 is the slope
- u is the error term or disturbance

Estimated regression equation

- The estimated regression equation shows how to calculate predicted values of the dependent variable using the values of the independent variable(s).

$$\hat{y} = b_0 + b_1 x_1 = x'b$$

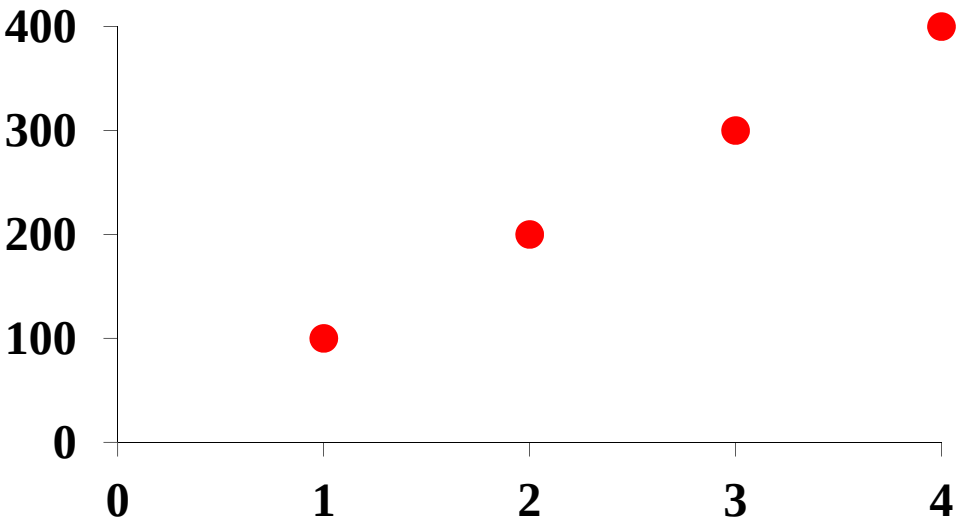
- Interpretation of the coefficients: one unit increase in *x* will increase the dependent variable *y* by b_1 units.
- Note that there is no error term when we predict the value of the depended variable.

- Regression residuals are calculated as the difference between the actual and predicted values of the dependent variable:

$$u = y - \hat{y} = y - b_0 - b_1 x_1 = y - x'b$$

Simple linear regression examples

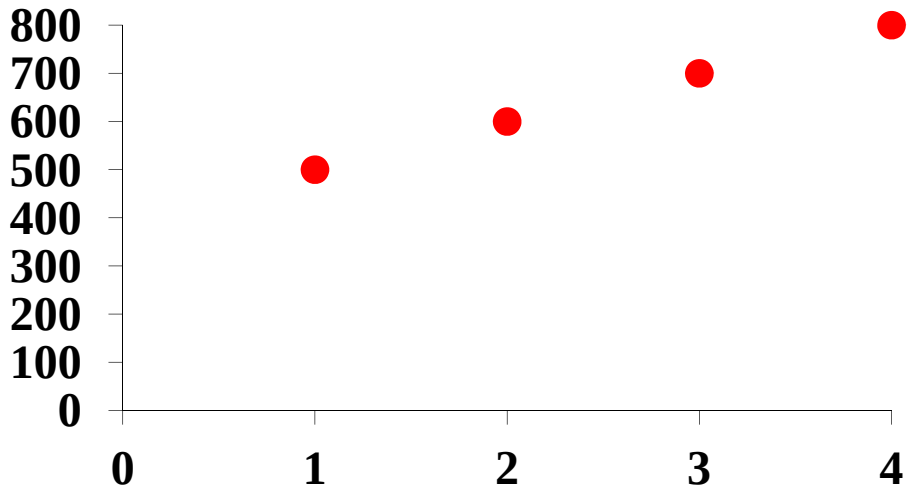
Regression line



x = number of credit cards
y = dollars spent
For each additional credit card, a person spends $100 more.

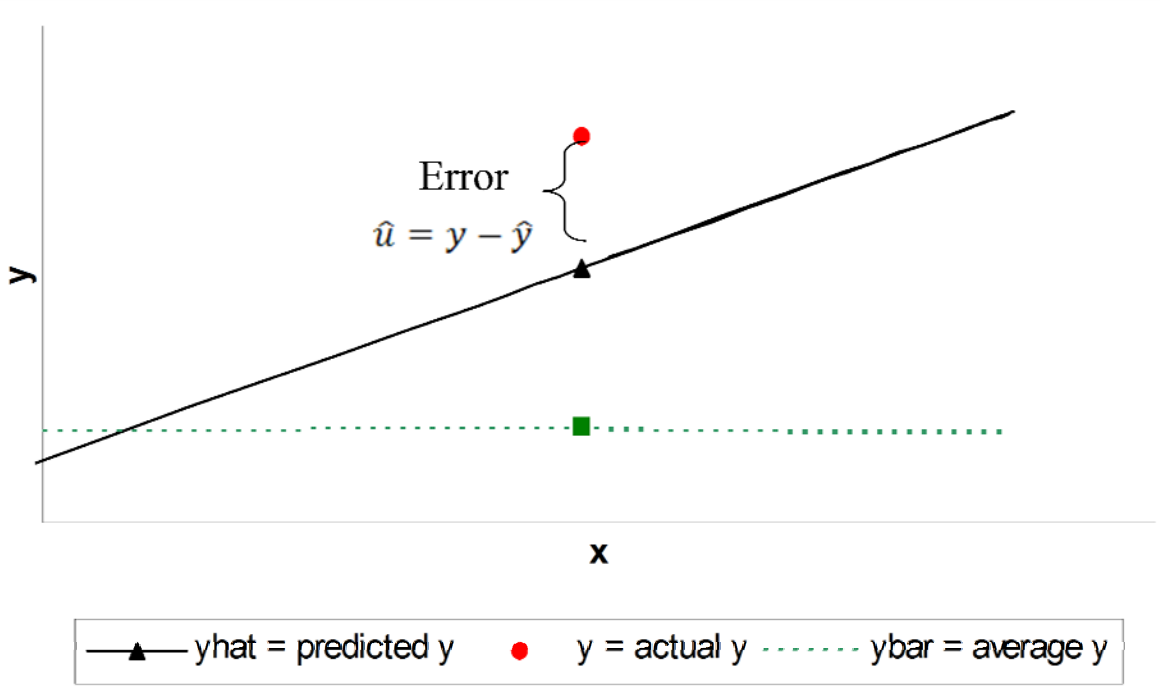
The equation for the line is $\hat{y} = b_0 + b_1 x_1 = 0 + 100x_1$
intercept = $b_0 = 0$ (when x_1=0, then $\hat{y} = b_0$)
slope = $b_1 = 100$ (when x_1 increases by 1, then $\hat{y}$ increases by b_1)

Regression line, new example with a positive intercept



The equation for the line is $\hat{y} = b_0 + b_1 x_1 = 400 + 100x_1$
intercept = b_0 = 400 (when x_1=0, then $\hat{y} = b_0$)
slope = b_1 = 100 (when x_1 increases by 1, then $\hat{y}$ increases by b_1)
For each additional credit card, a person spends $100 more.

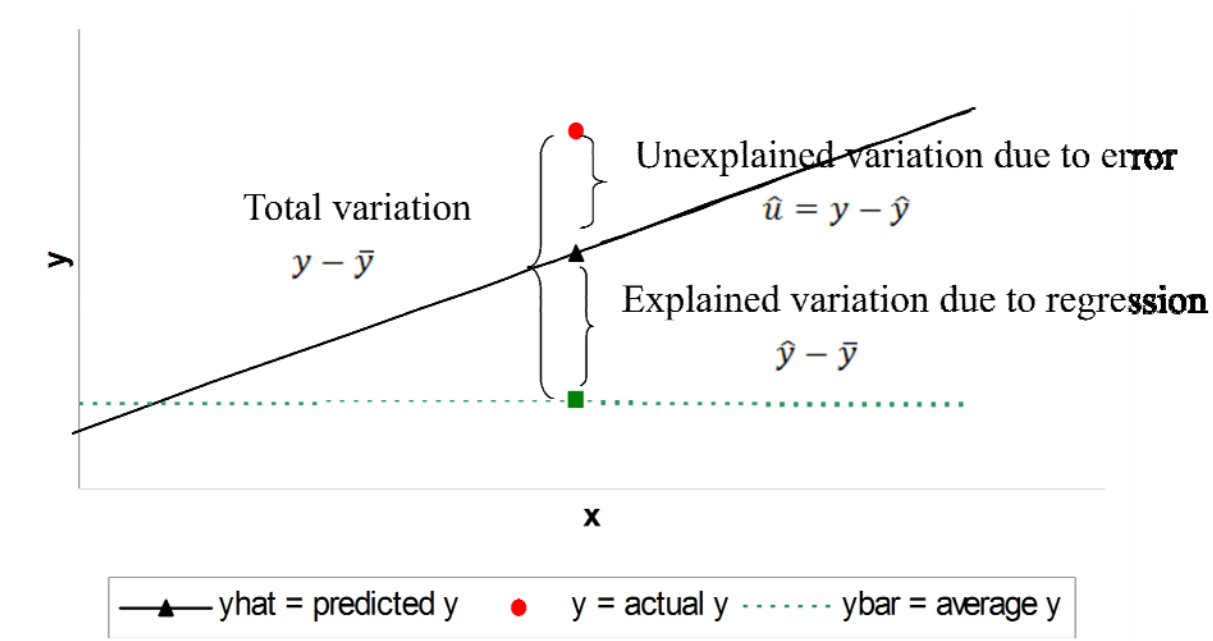
Regression error



The error is the difference between the actual values and the predicted values of the dependent variable:

$$u = y - \hat{y} = y - b_0 - b_1 x_1$$

Variations: total variation, explained variation and unexplained variation



$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

Total variation = explained variation due to regression + unexplained variation due to error
sum of squares total = sum of squares due to regression + sum of squares due to error
SST= SSR+SSE

The least squares method (OLS: ordinary least squares)

- The least squares method is used to calculate the coefficients so that the errors are as small as possible.
- We minimize the sum of squared residuals:

$$\sum u^2 = \sum (y - \hat{y})^2 = \sum (y - b_0 - b_1 x)^2$$

- In a simple linear regression the coefficients are calculated as:

$$b_1 = \frac{cov(x, y)}{var(x)}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

OLS regression in matrix form

- The regression line is specified as:

$$E(y|x) = x'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_P x_p$$

- Marginal effects in the linear regression model are the coefficients.

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j$$

- In multiple linear regression, the coefficients are calculated as:

$$b = \left(x'x\right)^{-1}(x'y)$$

- Assumptions of the OLS estimator:
  - Exogeneity of regressors
  - Homoscedasticity
  - Uncorrelated observations

Goodness of fit

R-squared

- The coefficient of determination (R-squared or R^2) provides a measure of the goodness of fit for the estimated regression equation.
- R^2 = SSR/SST = 1 – SSE/SST
- Values of R^2 close to 1 indicate perfect fit, values close to zero indicate poor fit.
- R^2 that is greater than 0.25 is considered good in the economics field.
- R-squared interpretation: if R-squared=0.8 then 80% of the variation is explained by the regression and the rest is due to error. So, we have a good fit.

Adjusted R-squared

- Problem: R^2 always increases when a new independent variable is added. This is because the SST is still the same but the SSE declines and SSR increases.
- Adjusted R-squared corrects for the number of independent variables and is preferred to R-squared.

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- where p is the number of independent variables, and n is the number of observations.

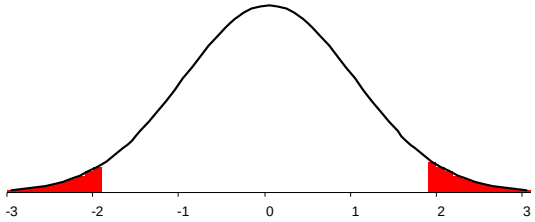
t-test for significance of one coefficient

- The t-test is used to determine whether the relationship between y and x_j is significant.

$$H_0: \beta_j = 0 \qquad\qquad H_a: \beta_j \neq 0$$

- The null hypothesis is that the coefficient is not significantly different than zero.
- The alternative hypothesis is that the coefficient is significantly different from zero.

- We use the t-distribution:
  - The test statistic t = coefficient /standard error
  - The critical values are from the t distribution
  - The test is a two-tailed test.



- Reject the null hypothesis and conclude that coefficient is significantly different from zero if:
  - The test statistic t is in the critical rejection zone
  - The p-value is less than 0.05
- The goal is to find coefficients that are significant.

F-test for overall significance of all coefficients

- Testing whether the relationship between y and all x variables is significant.
- The null hypothesis is that the coefficients are not jointly significantly different from zero.
- The alternative hypothesis is that the coefficients are jointly significantly different from zero.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a: \beta_1 \neq 0 \; or \; \beta_2 \neq 0 \; or \; \ldots \; \beta_p \neq 0$$

- Use the F-distribution
  - The test statistic F = MSR/MSE
  - The critical values are from the F distribution
  - The F-test is an upper one-tail test

ANOVA table

Total variation = explained variation due to regression + unexplained variation due to error

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic
Regression	SSR=$\sum(\hat{y}-\bar{y})^2$	p = number of independent variables	MSR=SSR/p	F=MSR/MSE
Error	SSE=$\sum(y-\hat{y})^2$	n-p-1	MSE=SSE/(n-p-1)	
Total	SST=$\sum(y-\bar{y})^2$	n-1 (n=number of observations)		

- Find critical values in the F table (significance level =0.05)
  - degrees of freedom in the numerator = number of independent variables = p
  - degrees of freedom in the denominator = n-p-1

- Reject the null hypothesis if the F-test statistic is greater than the F-critical value.
- Reject the null hypothesis if the p-value is less than 0.05.
- The goal is to find a regression model with coefficients that are jointly significant.