Bivariate Regression

Jerome Dumortier

Topics Covered

Motivation of regression analysis for policy

- Linking a single outcome of interest to one or more independent variables

Bivariate regression model components

Introduction

Goal of a regression model

- Estimation of the expected mean of the dependent variable given particular values of the independent variable(s)

Bivariate regression

- One dependent variable y and one independent variable x

Find the best linear relationship between y and x assuming each observation y_i is a function of x_i plus a random error term ϵ_i

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

We are looking for the conditional mean of Y given X, i.e., $E(Y|X)$

$$E(Y|X) = \beta_0 + \beta_1 \cdot X$$

Bivariate Regression Examples

Linking one variable to one outcome of interest even without (!) causal claims

- Relationship between years of education and earnings (outcome) to quantify returns to schooling
- Effect of government spending per student on student test scores
- Link between gasoline prices and (individual) vehicle miles traveled for transportation and climate policy
- Relationship between property tax rates housing prices (outcome) to inform local public finance
- Change in crime rate due to police presence or change in policy presence due to crime rate

Careful with causal claims due to “chicken-or-egg” situations

Causal Claims I

Education earnings

- Omitted variables such as ability, motivation, family background
- Reverse causality: Expected earnings influence schooling choices (endogeneity)

Spending per student and test scores

- Policy targeting: Low-performing districts receive more funding
- Omitted variables: parental inputs, school quality
- Simultaneity problem

Causal Claims II

Gas prices and vehicle miles traveled

- Omitted variables: Transit access, income, etc.
- Measurement issues: Price variation often reflects demand conditions

Property tax rates and housing prices

- Simultaneity: Tax base affects tax rates
- Sorting: Households choose locations based on taxes and amenities leading to an equilibrium (and not a causal) relationship

Police presence and crime

- Reverse causality: Police are deployed because crime is high

Multivariate Regression Examples

Linking multiple variables to one outcome of interest

- Effect of a minimum wage increase on employment, while accounting for local economic conditions, industry composition, and demographic structure
- Healthcare utilization as a function of insurance coverage, income, and age
- Relationship between air pollution exposure and health outcomes, while taking into account weather conditions and population density, and industry composition
- Quantifying household energy consumption as a function of income, energy prices, and housing characteristics to inform energy efficiency policy

Ordinary Least Square (OLS) Components

Example of the used car market

- Dependent variable: Price
- Independent variable: Miles

Every regression equation of the form $y = \beta_0 + \beta_1 \cdot x + \epsilon$ can be decomposed into four parts

- y: dependent variable
- x: independent variables
- β_0: intercept
- β_1: slope coefficient associated with the independent variable

The linear function does not tell us exactly what y will be for a given value of x but it does tell us the expected value of y, i.e., $E(y|x)$

OLS Setup I

Given a particular observation i, we have

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Given two values of β_0 and β_1, i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$, we can write

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i + \hat{\epsilon}_i$$

where ϵ_i represents the errors to obtain the observed y_i

- Theoretical difference between ϵ_i and $\hat{\epsilon}_i$ for which the former is the population parameter (always unknown) and the latter is the estimated error term given data

OLS Setup II

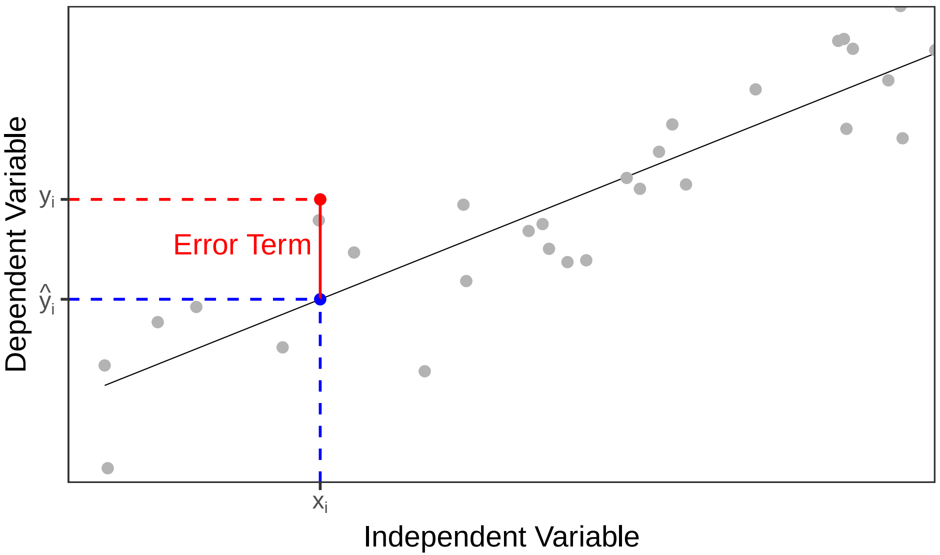
Rearranging leads to the following

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i$$

Minimization of the sum of the squared residuals:

$$\sum_{i=1}^{N} \hat{\epsilon}_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i\right)^2$$

OLS Setup Graphical Representation



OLS Optimal Solution

Equations necessary to solve the bivariate regression model

- Mean of x:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Mean of y:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

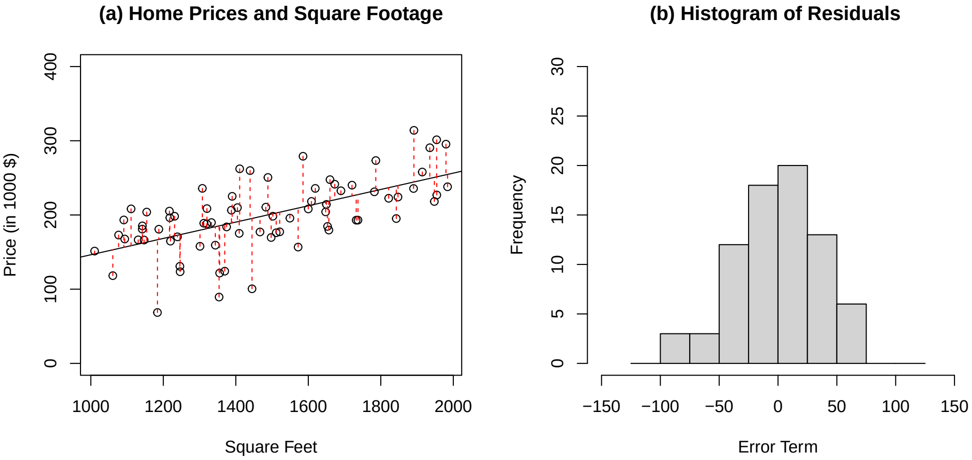
- Slope coefficients:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Intercept:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

Example: Home Values and Square Footage



Example: Used Cars

miles (x)	price (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
21	27	-15	6	-90	225
24	23	-12	2	-24	144
30	24	-6	3	-18	36
37	20	1	-1	-1	1
43	19	7	-2	-14	49
47	16	11	-5	-55	121
50	18	14	-3	-42	196

We have $\bar{x} = 36$ and $\bar{y} = 21$ as well as the following:

$$\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = -244 \quad \text{and} \quad \sum_{i=1}^{N} (x_i - \bar{x})^2 = 772$$

R^2: Measuring the Strength of the Relationship I

Goodness of fit measure decomposes the variation of Y into two components, i.e., the (1) unexplained variation and the (2) explained variation: $R^2 \in [0, 1]$.

Unexplained or residual variation

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Explained variation

$$ESS = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

Total variation:

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

R^2: Measuring the Strength of the Relationship II

R^2 as the proportion of the total variation in Y explained by independent variables. Note that since $TSS = RSS + ESS$:

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

R^2 defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Adjusted R^2 (for the case of multiple independent variables) where k is the number of variables:

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k}$$

Hypothesis Testing I

Standard error for the slope coefficient:

$$se(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

Standard error for the intercept:

$$se(\hat{\beta}_0) = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{n\sum_{i=1}^{N}(x_i - \bar{x})^2}}\sigma$$

Estimate for the variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{n-2}$$

Hypothesis Testing II

Determination of statistical significance between variables:

- Assumption of normally distributed error terms
- t-statistic with $n-2$ degrees of freedom

Specific hypothesis tests are H_0: $\beta_0 = 0$ and H_0: $\beta_1 = 0$. The test statistic for β_i can be written as

$$\frac{\hat{\beta}_i - \beta_i}{se_{\hat{\beta}_i}} \sim t_{n-2}$$

The hypothesis test is never conducted manually and every statistical software conducts and reports the results of the hypothesis test.

Numerical Example: Post Estimation

miles (x)	price (y)	x_i^2	$\hat{y}$	e_i	e_i^2	$(y_i - \bar{y})^2$
21	27	441	25.74	1.26	1.59	36
24	23	576	24.79	-1.79	3.21	4
30	24	900	22.90	1.10	1.22	9
37	20	1369	20.68	-0.68	0.47	1
43	19	1849	18.79	0.21	0.05	4
47	16	2209	17.52	-1.52	2.32	25
50	18	2500	16.58	1.42	2.03	9

Note that $\sum e_i^2 = 10.89$, $\sum x_i^2 = 10.89$, and $\sum (y_i - \bar{y})^2 = 88$

Numerical Example: R^2 and Standard Errors

Goodness of fit R^2:

$$R^2 = 1 - \frac{10.89}{88} = 0.876$$

For the standard errors, we have $\hat{\sigma} = \sqrt{10.89/5} = 1.476$ and thus,

$$se(\hat{\beta}_0) = \sqrt{\frac{9844}{7 \cdot 772}} \cdot 1.476 = 1.99$$

$$se(\hat{\beta}_1) = \frac{1.476}{\sqrt{772}} = 0.053$$

Adjusted R^2:

$$\bar{R}^2 = 1 - (1 - 0.876) \cdot 6/5 = 0.8512$$

The manual calculations match the output from R.

Assumptions

Important assumptions for unbiasedness of the coefficient estimates:

- A1: Linear regression model, i.e., linear in terms of coefficients
- A2: Zero mean value of error terms ϵ, i.e., $E(\epsilon_i|x_i) = 0$
- A3: Homoscedasticity or equal variance of ϵ_i, i.e., $Var(\epsilon_i) = \sigma^2$
- A4: No autocorrelation between the error terms, i.e., $Cov(\epsilon_i, \epsilon_j) = 0$
- A5: No covariance between ϵ_i and x_i
- A6: Number of observations is greater than number of parameters to be estimated
- A7: No multicollinearity

A1: Linear Regression Model

Regression model that is linear in parameters:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon$$

Note that the following models are also linear in parameters:

$$y_i = \beta_0 + \beta_1 \cdot x_i^2 + \epsilon$$

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \epsilon$$

The following model is linear in parameters:

$$y = e^{\beta_0 + \beta_1 \cdot x_i}$$

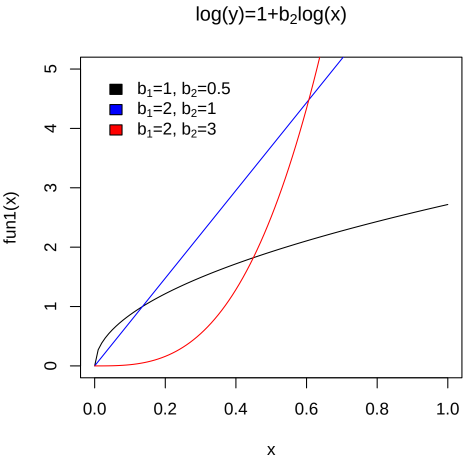
The last model can be estimated by taking the natural logarithm of both sides.

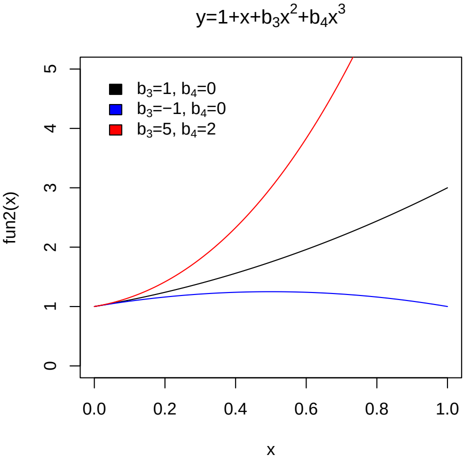
A1: Linear Regression Model

Despite the fact that the regression model is linear, non-linear relationships can be measured:

- Relation between consumption and income might be non linear since a change in consumption due to extra income may decrease with income.
- Relationship between income and education can exhibit a non-linear form because a change in income due to more education may decrease with more education.

Example of Linear Regression Models





A2: Zero Mean Value of Error Terms

Expected value of the error term is 0:

$$E(\epsilon_i|X_i) = 0 \quad \text{for all } i$$

See the histogram of residuals a couple of slides back.

A3: Homoscedasticity

The variance of the error terms is constant:

$$Var(\epsilon_i|x_i) = E(\epsilon_i^2|x_i) = \sigma^2$$

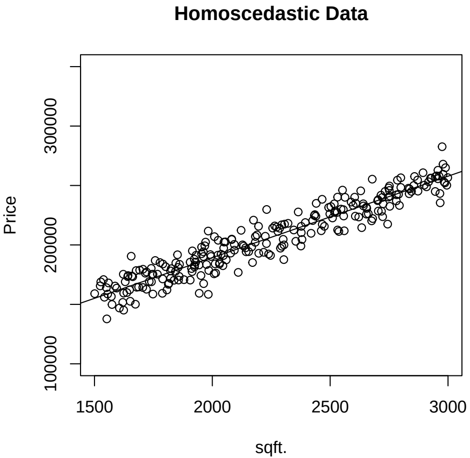
The assumption of constant variance is known as homoscedasticity. A violation of this assumption represents heteroscedasticity. Consider the following examples:

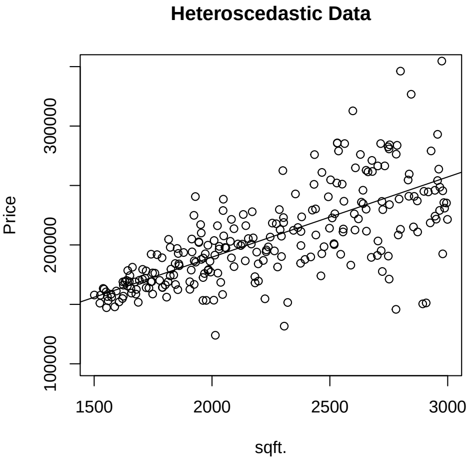
- Weekly consumption expenditures increases with income but the variability is higher with high-income families.

Consequences:

- No consequence on coefficient estimates
- Inflated standard errors

A3: Homoscedasticity vs. Heteroscedasticity





A4-A7: Other Assumptions

A4: No autocorrelation between the disturbance terms

$$E(\epsilon_i \epsilon_j) = 0 \quad \text{for all } i \neq j$$

A5: No covariance between ϵ_i and x_i

$$Cov(\epsilon_i, X_i) = 0$$

A6: Full rank:

- More observations than variables to be estimated
- Analogy: You cannot solve for three unknowns with two equations

A7: Multicollinearity:

- Near perfect linear relationships between independent variables should be avoided

Application in R

```
##
## Call:
## lm(formula = price ~ miles, data = cars)
##
## Residuals:
##       1       2       3       4       5       6       7
##  1.2591 -1.7927  1.1036 -0.6839  0.2124 -1.5233  1.4249
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.37824    1.99101  16.262  1.6e-05 ***
## miles       -0.31606    0.05309  -5.953  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.475 on 5 degrees of freedom
## Multiple R-squared:  0.8764, Adjusted R-squared:  0.8516
## F-statistic: 35.44 on 1 and 5 DF,  p-value: 0.001912
```

Used Car Market: R/RStudio Output

```
bhat = lm(price~miles,data=honda)
summary(bhat)
```

```
##
## Call:
## lm(formula = price ~ miles, data = honda)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2453.6 -1055.3  -139.0   604.2  5389.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.205e+04  4.890e+02  45.095  < 2e-16 ***
## miles       -6.501e-02  1.251e-02  -5.198 1.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1455 on 79 degrees of freedom
## Multiple R-squared:  0.2549, Adjusted R-squared:  0.2454
## F-statistic: 27.02 on 1 and 79 DF,  p-value: 1.539e-06
```

CLRM: Assumptions

Required assumptions for unbiasedness coefficient estimates:

- A1: Linear in terms of coefficients, i.e., $y = \beta_0 + \beta_1 \cdot x$
- A2: Zero mean value of error terms ϵ, i.e., $E(\epsilon_i | x_i) = 0$
- A3: Homoscedasticity (equal variance) of ϵ_i, i.e., $Var(\epsilon_i) = \sigma^2$
- A4: No autocorrelation between the error terms, i.e., $Cov(\epsilon_i, \epsilon_j) = 0$
- A5: No covariance between ϵ_i and x_i
- A6: Number of observations is greater than number of parameters to be estimated
- A7: Full rank and absence of (perfect) multicollinearity

CLRM: Relaxing Assumptions

Relaxing the assumptions of the classical regression model requires regression diagnostics and/or different regression approaches

- Multicollinearity: Correlation between independent variables are correlated with each other?
  - Beds and bathrooms in a home value model
  - Multicollinearity occurs between two or more (!) independent variables
- Heteroscedasticity: Errors variance not constant
- Autocorrelation between error terms
- Inclusion of irrelevant or exclusion of relevant independent variables