Linear Regression in R

Data Avicenna

Department of Economics
Universitas Gadjah Mada

May 28, 2025

Outline

What is Regression Analysis?

- Regression analysis is a statistical method that examines the relationship between a dependent variable (Y) and one or more independent variables (X)
- Linear regression specifically models linear relationships between these variables
- **Applications**:
  - **Causal inference**: using data to estimate the effect on an outcome of interest of an intervention that changes the value of another variable.
  - **Prediction**: using the observed value of some variable to predict the value of another variable.
- R provides powerful tools for regression analysis through both base R functions and specialized packages

Simple and Multiple Linear Regression

- **Simple Linear Regression**: One independent variable (X)
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
- **Multiple Linear Regression**: Two or more independent variables
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$
- Today's focus:
  - Understanding the theoretical basis
  - Implementing both types in R
  - Interpreting results and diagnostics
  - Addressing common issues

Simple Linear Regression Model

- The simplest regression model includes a dependent variable (Y) and one independent variable (X)
- **Mathematical representation:**

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

Where:

- Y_i is the dependent variable (outcome)
- X_i is the independent variable (predictor)
- β_0 is the intercept (estimated value of Y when X = 0)
- β_1 is the slope (effect of X on Y)
- u_i is the error term (unexplained variation)

Ordinary Least Squares (OLS)

- Main idea of OLS: The OLS estimator chooses the regression coefficients such that the estimated regression line is as "close" as possible to the observed data points.
- Closeness is measured by the sum of squared mistakes in predicting Y given X. Let b_0 and b_1 be some estimators of β_0 and β_1. OLS minimizes the sum of squared mistakes:

$$\hat{\beta}_0, \hat{\beta}_1 := \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- OLS estimators for simple regression:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{Cov(X, Y)}{Var(X)} \quad (2)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (3)$$

Ordinary Least Squares (OLS)

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$
$$\hat{u}_i = Y_i - \hat{Y}_i.$$

The estimated intercept $\hat{\beta}_0$, the slope parameter $\hat{\beta}_1$ and the residuals ($\hat{u}_i$) 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).

Ordinary Least Squares (OLS)

Key points:

- The OLS estimator of the slope is equal to the ratio of sample covariance and sample variance!
- The OLS estimators are functions of the sample data only
- Given the sample data (X_i, Y_i) we can compute the $\hat{\beta}_1$ and then $\hat{\beta}_0$
- Computer programs such as Stata and R easily calculate $\hat{\beta}_1$ and $\hat{\beta}_0$ for you

Implementing Simple Linear Regression in R

We will be using the AER package from Applied Econometrics with R (Kleiber and Zeilis, 2008). The `lm()` function in R is used for fitting linear models:

```
# install.packages("AER")
# Load necessary libraries
library(AER)
library(ggplot2)  # For visualization

# Load the CASchools dataset
data(CASchools)

# Create student-teacher ratio and test score variables
CASchools$STR <- CASchools$students/CASchools$teachers
CASchools$score <- (CASchools$read + CASchools$math)/2

# Fit a simple linear regression model
SLR <- lm(score ~ STR, data = CASchools)

# View a summary of the model
summary(SLR)
```

Visualizing the Simple Regression

```r
# Basic scatter plot with regression line
plot(CASchools$STR, CASchools$score,
     main = "Test Score vs. Student-Teacher Ratio",
     xlab = "Student-Teacher Ratio (STR)",
     ylab = "Test Score",
     pch = 20, col = "steelblue")

# Add the regression line
abline(model, col = "red", lwd = 2)

# Alternative using ggplot2
ggplot(CASchools, aes(x = STR, y = score)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Test Score vs. Student-Teacher Ratio",
       x = "Student-Teacher Ratio (STR)",
       y = "Test Score")
```

Understanding the Output

```r
# Store the model summary
model_summary <- summary(SLR)

# Coefficients
model_summary$coefficients

# Extracting specific values
beta_0 <- coef(SLR)[1]  # Intercept
beta_1 <- coef(SLR)[2]  # Slope
std_errors <- coef(summary(SLR))[, 2]  # Standard errors
t_values <- coef(summary(SLR))[, 3]  # t-values
p_values <- coef(summary(SLR))[, 4]  # p-values

# Model fit statistics
r_squared <- model_summary$r.squared
adj_r_squared <- model_summary$adj.r.squared
residual_se <- model_summary$sigma
```

Interpreting the Simple Regression Results

- **Coefficients**:
  - $\hat{\beta}_0$ (Intercept): Estimated average test score when STR $= 0$
  - $\hat{\beta}_1$ (Slope): Estimated change in test score associated with a one-unit increase in STR
- **Statistical significance**:
  - t-values and p-values help determine if coefficients are statistically significant
  - Typically use significance levels of 0.05 or 0.01
- **Goodness of fit**:
  - R^2: Proportion of variance in Y explained by X
  - Adjusted R^2: R^2 adjusted for the number of predictors
  - Residual standard error: Estimate of the standard deviation of the error term

Assumptions of Linear Regression

For a valid causal inference, OLS assumes:

1. Conditional mean independence (CMI): $E(u_i|X_i) = 0$
2. Sample data are i.i.d. draw from population distribution
3. Large outliers are unlikely

Violation of these assumptions can lead to biased coefficient estimates.

If the three least squares assumptions hold and if the error is homoskedastic ($Var(u_i|X_i)$ is constant), the OLS estimator is **Best Linear Unbiased Estimator (BLUE)**.

Multiple Linear Regression Model

- Extends simple regression to include multiple predictors
- **Mathematical representation:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i \tag{4}$$

- Advantage: Addresses omitted variable bias

Omitted Variable Bias

- Occurs when an important predictor is excluded from the model
- Two conditions for omitted variable bias:
  1. The regressor, X, is correlated with the omitted variable
  2. The omitted variable is a determinant of the dependent variable (Y)
- Results in biased coefficient estimates

Omitted Variable Bias

Example:

- In the test score model, excluding the percentage of English learners in the school district could bias the estimated effect of student-teacher ratio.
- It is plausible that the ability to speak, read and write English is an important factor for successful learning.
- Also, it is conceivable that the share of English learning students is bigger in school districts where class sizes are relatively large: think of poor urban districts where a lot of immigrants live.

Implementing Multiple Regression in R

We call potentially omitted variables that are included in the regression model as **control** variables. They control for factors that are correlated with the explanatory variable of interest (X) and may influence the outcome (Y).

Let's add the percentage of English learners as a **control** variable:

```r
# Multiple regression with STR and english
MLR <- lm(score ~ STR + english, data = CASchools)

# Summary of the model
summary(MLR)

# Compare with simple regression
summary(SLR)
```

Adding More Control Variables

Now, let's add another control variable: **lunch**. This represents the percentage of students that qualify for a free or subsidized lunch in school due to family incomes below a certain threshold.

An argument to include this variable is that students' economic background are strongly related to outside learning opportunities: think of wealthy parents that are able to provide time and/or money for private tuition of their children.

Adding More Control Variables

```
# Add more variables to the model
model_expanded <- lm(score ~ STR + english + lunch,
data = CASchools)

summary(model_expanded)

Call:
lm(formula = score ~ STR + english + lunch, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-32.849  -5.151  -0.308   5.243  31.501

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 700.14996    4.68569 149.423  < 2e-16 ***
STR          -0.99831    0.23875  -4.181 3.54e-05 ***
english      -0.12157    0.03232  -3.762 0.000193 ***
lunch        -0.54735    0.02160 -25.341  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding More Control Variables

Thus, the estimated regression line is

$$\widehat{TestScore} = \underset{(5.56)}{700.15} - \underset{(0.27)}{1.00} \times STR - \underset{(0.03)}{0.12} \times english - \underset{(0.02)}{0.55} \times lunch. \tag{5}$$

We observe no substantial changes in the conclusion about the effect of STR on $TestScore$: the coefficient on STR changes by only 0.1 and retains its significance.

Although the difference in estimated coefficients is not big in this case, it is useful to keep *lunch* to make the assumption of conditional mean independence more credible (see Chapter 7.5 of Stock and Watson).

Interpreting Multiple Regression Coefficients

- In multiple regression, each coefficient is interpreted "holding all other variables constant"
- $\hat{\beta}_1$: Expected change in Y associated with a one-unit increase in X_1, holding all other variables constant
- This is often called the "ceteris paribus" interpretation
- Comparing coefficients across models can reveal potential omitted variable bias
- Changes in coefficient magnitude or significance when adding variables can provide insights into relationships between predictors

Categorical Variables in Regression

```r
# Create a dummy variable for high STR
CASchools$high_STR <- ifelse(CASchools$STR > median(CASchools$STR), 1, 0)

# Regression with dummy variable
model_dummy <- lm(score ~ high_STR, data = CASchools)
summary(model_dummy)
```

Creating Publication-Quality Regression Tables

```r
# Using stargazer for publication-quality tables
library(stargazer)
models <- list(SLR, MLR, model_expanded)

stargazer(models,
          title = "Test Score Regression Results",
          column.labels = c("Simple", "With English", "Full Model"),
          covariate.labels = c("Student-Teacher Ratio",
                               "% English Learners", "% Free Lunch"),
          dep.var.labels = "Test Score",
          type = "html",
          out = "table.html",           # Open in browser, copy to Word
          digits = 3,
          star.cutoffs = c(0.05, 0.01, 0.001),
          omit.stat = c("f", "ser"))
```

Creating Publication-Quality Regression Tables

- You can copy and paste your table to Word, so that it becomes editable.

Test Score Regression Results

	Dependent variable:		
	Test Score		
	Simple	With English	Full Model
	(1)	(2)	(3)
Student-Teacher Ratio	-2.280***	-1.101**	-0.998***
	(0.480)	(0.380)	(0.239)
% English Learners		-0.650***	-0.122***
		(0.039)	(0.032)
% Free Lunch			-0.547***
			(0.022)
Constant	698.933***	686.032***	700.150***
	(9.467)	(7.411)	(4.686)
Observations	420	420	420
R^2	0.051	0.426	0.775
Adjusted R^2	0.049	0.424	0.773
Note:	*p<0.05; **p<0.01; ***p<0.001		

References

- Hanck, C., Arnold, M., Gerber, A., Schmelzer, M. (n.d.). Introduction to Econometrics with R. Retrieved April 24, 2025, from https://www.econometrics-with-r.org/
- Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
- Stock, J. H., & Watson, M. W. (2020). Introduction to econometrics (4th edition).
- Lecture notes for EMET6010 Applied Macro and Financial Econometrics, by Thomas Tao Yang (Research School of Economics, The Australian National University).