Statistical Theory II
Practical Labs

Lab 5: Linear Regression

Paulius Kazlauskas

Plan for today

1. Introduction to linear regression
2. Population vs. sample & OLS
3. Goodness of fit and interpreting results
4. Extension: multiple regression
5. Examples in R
6. Lab work & solutions

Why linear regression?

- One of the most widely used statistical techniques in economics and the social sciences
- Core tool for empirical research — you will encounter it in nearly every applied paper
- Foundational for more advanced methods (IV, panel data, difference-in-differences . . .)
- Valued by employers in both academia and policy institutions

Today's goal: understand what regression does, how to run it in R, and how to read and interpret the output.

What is regression?

Regression describes the relationship between one or more **independent variables** and the **expected value** of a dependent variable:

$$y = f(x_1, x_2, \ldots, x_n)$$

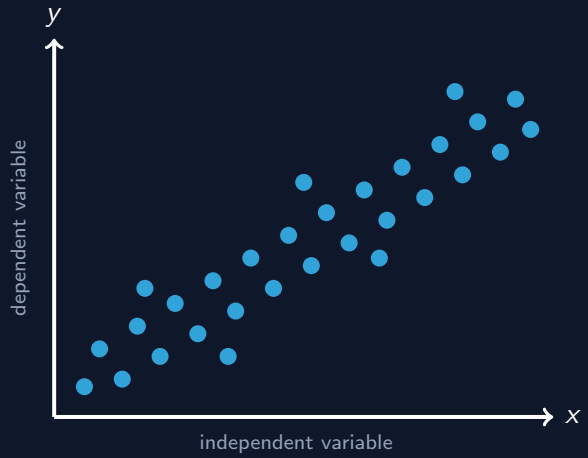
Linear pairwise regression is the simplest case — one predictor, linear relationship:

Intuition

Given a cloud of data points, we want to find the straight line that fits them *as well as possible* — and then use that line to understand and predict y from x.

Let's see what this looks like visually →

What does it look like? — The raw data

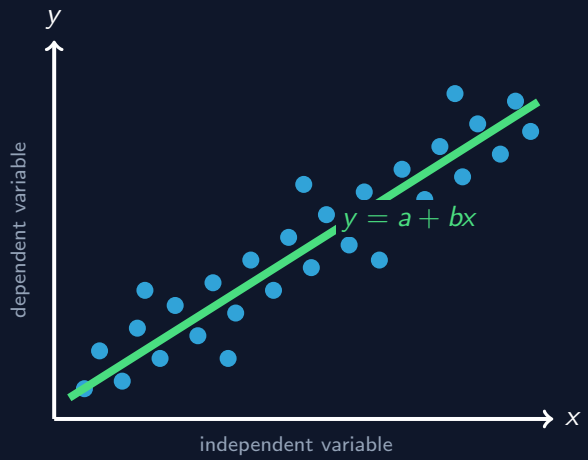


We have data — but no model yet.

There seems to be a pattern: as x increases, y tends to increase too.

How do we capture this pattern with a line?

Finding the best-fitting line



We want to find the line:

$$y = a + bx$$

that best fits the data.

- a is the **intercept** — where the line crosses the y-axis
- b is the **slope** — how much y changes when x increases by one unit

But which line is the "best"? We need a criterion — that comes next.

The pairwise regression model

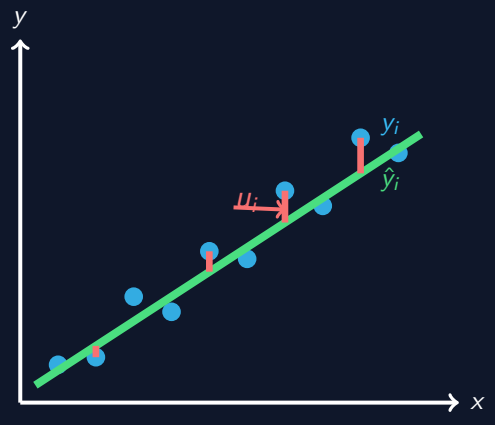
The formal model is:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- y_i dependent (outcome) variable
- x_i independent (predictor) variable
- β_0 **intercept** — expected value of y when $x = 0$
- β_1 **slope** — change in y for a one-unit increase in x
- u_i **error** — the part of y_i not explained by the line

The error u_i absorbs everything the model does not account for: omitted variables, measurement error, and genuine randomness.

The error term — visually



The **error** u_i is the vertical gap between the observed point y_i and the line:

$$u_i = y_i - (a + b\,x_i)$$

The line $y = a + bx$ is chosen so that these errors are collectively as small as possible.

Some errors are positive (point above the line), some negative (point below). A good fit means these are small overall.

Population vs sample

Population — true, unknown

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The true relationship that exists in the world. We **never** observe β_0 and β_1 directly.

Sample — estimated from data

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

Our best guess at β_0 and β_1 using observed data. The hat $\hat{}$ denotes an estimate.

The estimated residuals are therefore:

$$\hat{u}_i \;=\; y_i - \hat{y}_i \;=\; y_i - \hat{\beta}_0 - \hat{\beta}_1\, x_i$$

The goal of OLS is to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that these residuals are, in total, as small as possible.

Ordinary Least Squares (OLS)

We choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to **minimise the sum of squared residuals**:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2$$

Solving this minimisation problem gives closed-form formulas:

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(y_i, x_i)}{\mathrm{Var}(x_i)} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Why squared errors? Squaring ensures positive and negative residuals do not cancel, and penalises large deviations more heavily than small ones.

OLS assumptions & the Gauss-Markov theorem

For OLS to produce reliable estimates, the following conditions should hold:

- y is a linear function of x and u
- The sample is representative of the population
- x_i is uncorrelated with the errors: $\mathrm{E}[u_i] = 0$
- In *multiple regression*: no perfect multicollinearity between predictors
- Errors are homoscedastic (constant variance) and not autocorrelated
- Errors follow a normal distribution

Gauss-Markov theorem

When these conditions hold, OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ are **BLUE** — Best Linear Unbiased Estimators. They are unbiased ($\mathrm{E}[\hat{\beta}] = \beta$) and have the lowest variance of all linear unbiased estimators.

Interpreting the coefficients — Example 1

Suppose we estimate a regression showing how students' grades depend on study hours:

Study hours and grades

$$\widehat{\text{Grade}}_i = 3.5 + 0.2 \cdot \text{Study}_i + \hat{u}_i$$

- $\hat{\beta}_0 = 3.5$: a student who does *zero* study hours is expected to score **3.5**
- $\hat{\beta}_1 = 0.2$: each additional hour of study raises the predicted grade by **0.2 points**, on average
- $\hat{u}_i$: the residual — whatever makes student i's actual grade differ from the prediction (talent, luck, exam difficulty, ...)

The slope $\hat{\beta}_1$ is usually what we care about most: the estimated effect of x on y.

Interpreting the coefficients — Example 2

Now suppose we estimate how salary depends on years of experience:

Experience and salary

$$\widehat{\text{Salary}}_i = 730 + 180 \cdot \text{Experience}_i + \hat{u}_i$$

- $\hat{\beta}_0 = 730$: the predicted salary for someone with *zero* years of experience is **\$730**
- $\hat{\beta}_1 = 180$: each additional year of experience is associated with **\$180 more** in predicted salary, on average
- $\hat{u}_i$: captures everything else affecting salary — education, sector, negotiation skills, ...

The intercept is often not directly meaningful (zero experience may be outside the data range). Focus on the slope.

How good is our model? — R^2

We decompose the total variation in Y into explained and unexplained parts:

SST $= \sum_{i=1}^{n} (y_i - \bar{y})^2$ Total variation in Y

SSE $= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ Variation *explained* by the model

SSR $= \sum_{i=1}^{n} \hat{u}_i^2$ Variation *left unexplained* (residuals)

Note: $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$

$$R^2 = \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} \qquad R^2 \in [0, 1]$$

Interpreting R^2

- $R^2 = 1$: the model explains *all* variation in Y — every point lies exactly on the line
- $R^2 = 0$: the model explains *none* of the variation — x tells us nothing about y
- In practice: 0.5–0.7 is common in social science; what is "good" depends heavily on context

A caution for multiple regression: Adjusted R^2

R^2 **always increases** when you add a predictor — even a useless one.

The **Adjusted** R^2 penalises for model complexity:

$$\bar{R}^2 = 1 - \frac{SSR\,/\,(n-k-1)}{SST\,/\,(n-1)}$$

where k is the number of predictors. Use $\bar{R}^2$ when comparing models of different sizes.

Are the coefficients statistically significant?

Estimating $\hat{\beta}_k$ is not enough — we also need to ask: **could this result just be noise?**

For each coefficient, R automatically runs a *t*-**test**:

$$H_0 : \beta_k = 0 \qquad \text{vs} \qquad H_1 : \beta_k \neq 0$$

The test statistic is:

$$t = \frac{\hat{\beta}_k}{\mathrm{SE}(\hat{\beta}_k)}$$

$\hat{\beta}_k$	our estimated coefficient
$\mathrm{SE}(\hat{\beta}_k)$	its standard error — measures estimation uncertainty
t	how many standard errors away from zero our estimate is

Reading significance in the R output

Rule of thumb

If $|t| \gtrsim 2$, we reject H_0 at the 5% significance level and conclude the coefficient is **statistically different from zero**.

R reports a p-value — the probability of observing a t-statistic at least this large *if* H_0 *were true*. Stars give you a quick read:

Stars	p-value	Interpretation
***	< 0.001	Very strong evidence against H_0
**	< 0.01	Strong evidence
*	< 0.05	Moderate evidence — conventional threshold
.	< 0.10	Weak evidence
(none)	≥ 0.10	No significant evidence against H_0

Extending to multiple predictors

Outcomes usually depend on more than one variable. Multiple regression extends the pairwise model naturally:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

Each $\hat{\beta}_j$ is a **partial effect**: the change in y for a one-unit increase in x_j, *holding all other predictors fixed (ceteris paribus)*.

Today's dataset — Real Estate Valuation

$$\widehat{\text{Price}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i + \hat{\beta}_2 \cdot \text{MRT distance}_i + \hat{\beta}_3 \cdot \text{Stores}_i + \hat{u}_i$$

Each coefficient captures the *separate* contribution of that variable, after accounting for the others.

Why multiple regression matters — omitted variable bias

If a variable that affects Y is left out of the model, its effect gets *absorbed* into the remaining coefficients — biasing all estimates.

Example from our dataset

Regressing house price on *age only* gives a misleading coefficient, because older houses also tend to be further from MRT stations — the two are correlated. Including both simultaneously isolates the *true* partial effect of age from the effect of distance.

Multiple regression is how we *control* for confounders — the foundation of most empirical work in economics.

Running the regression in R

```r
library(readxl)

# Read in the data
data_RE <- read_excel( Real estate valuation data set.xlsx )

# Fit the multiple regression model
model_RE <- lm(
  `Y house price of unit area` ~
    `X2 house age` +
    `X3 distance to the nearest MRT station` +
    `X4 number of convenience stores`,
  data = data_RE
)

# Display the results
summary(model_RE)
```

Reading the `summary()` output

Coefficients table

	Estimate	Std. Error	t-value	Pr($>\|t\|$)
(Intercept)	45.11	2.31	19.5	<2e-16 ***
X2 house age	−0.27	0.05	−5.1	5e-07 ***
X3 MRT distance	−0.005	0.001	−8.2	<2e-16 ***
X4 stores	1.31	0.18	7.2	2e-12 ***

Multiple R^2: **0.576** Adjusted R^2: **0.572** F-stat: **186.0**, $p <$ 2.2e-16

All three predictors are significant at $p < 0.001$. The model explains 57.6% of the variation in house prices.

Predicted values and scatter plot

```r
# Get fitted (predicted) values from the model
data_RE$predicted_price <- predict(model_RE)

# Base R scatter: house age vs fitted price
plot(data_RE$`X2 house age`,
     data_RE$predicted_price,
     main =  House Age vs Fitted Price ,
     xlab =  House age (years) ,
     ylab =  Fitted price (per unit area) ,
     pch  = 19, col =  lightblue )

# Add a regression line
abline(lm(predicted_price ~ `X2 house age`,
          data = data_RE),
       col =  orange , lwd = 2)
```

Model diagnostics — a quick check

After fitting your model, it is good practice to run two quick visual checks:

```
plot(model_RE, which = 1)   # Residuals vs Fitted
plot(model_RE, which = 2)   # Normal Q-Q
```

Residuals vs Fitted

Plots each residual $\hat{u}_i$ against its fitted value $\hat{y}_i$. You want random scatter around the horizontal zero line — no curves, no funnels. A clear pattern here means the model is missing something.

Normal Q-Q

Checks whether the residuals are approximately normally distributed. Points should lie close to the diagonal line. Systematic departures (an S-shape or heavy tails) suggest the normality assumption may not hold.

These two plots are sufficient for this course. The goal is to look at them, describe