The Linear Regression Model

Bas Machielsen

2025-06-12

What is Econometrics?

Econometrics is the use of statistical methods to:

1. **Estimate** economic relationships.
2. **Test** economic theories.
3. **Evaluate** and implement government and business policy.
4. **Forecast** economic variables.

It's where economic theory meets real-world data. Theory proposes relationships (e.g., Law of Demand), but econometrics tells us the magnitude and statistical significance of these relationships.

Why study econometrics?

- It allows you to **quantify** the relationships that you learn about in your other economics courses.
    - *By how much* does demand fall if we raise the price by 10%?
    - What is the effect of an additional year of education on future wages?
    - It helps distinguish between **correlation** and **causation**.
    - It is an essential tool for empirical research in economics and finance, and a highly valued skill in the job market.

The Nature of Economic Data

The type of data we have determines the econometric methods we should use.

- **Cross-Sectional Data:** A snapshot of many different individuals, households, firms, countries, etc., at a *single point in time*.
  - *Example:* A survey of 500 individuals in 2023, with data on their wage, education, gender, and age.
- **Time Series Data:** Observations on a single entity (e.g., a country, a company) collected over *multiple time periods*.
  - *Example:* Data on U.S. GDP, inflation, and unemployment from 1950 to 2023.
- **Pooled Cross-Sections:** A combination of two or more cross-sectional datasets from different time periods. The individuals are different in each period.
  - *Example:* A random survey of households in 1990, and another *different* random survey of households in 2020.
- **Panel (or Longitudinal) Data:** The *same* cross-sectional units are followed over time.
  - *Example:* Tracking the wage, education, and city of residence for the same 500 individuals every year from 2010 to 2020.

The Concept of a Model

The Population Regression Function (PRF)

Let's say we are interested in the relationship between wages (y) and years of education (x). Economic theory suggests a positive relationship.

We can model the *average* wage for a given level of education. This is the **Population Regression Function (PRF)**:

$$E(y|x) = \beta_0 + \beta_1 x$$

- $E(y|x)$ is the **expected value (average) of y, given a value of x**.
- β_0 (beta-nought) is the **population intercept**.
- β_1 (beta-one) is the **population slope**. These are unknown constants (parameters) that we want to estimate.

The PRF represents the true, but unknown, relationship in the population.

The Stochastic Error Term

Of course, not everyone with the same level of education has the same wage. Other factors matter (experience, innate ability, location, luck, etc.).

We capture all these other unobserved factors in a **stochastic error term**, u.

Our individual-level population model is:

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

Where: * y_i is the wage of individual i. * x_i is the education of individual i. * u_i is the error term for individual i. It represents the deviation of individual i's actual wage from the population average, $E(y|x_i)$.

By definition of the conditional expectation, $E(u|x) = 0$. The average of the unobserved factors does not depend on the level of education.

2. The Sample Regression Function (SRF)

From Population to Sample

We can't observe the entire population. We only have a sample of data. Our goal is to use the sample data to *estimate* the unknown population parameters β_0 and β_1.

The **Sample Regression Function (SRF)** is our estimate of the PRF:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{y}$ (y-hat) is the **predicted** or **fitted** value of y.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the **estimators** of β_0 and β_1. They are statistics calculated from our sample data.

```
ggplot(dat, aes(x=educ, y=wage)) +
  geom_point(alpha=0.6) +
  labs(title="Sample Data and a Potential Regression Line", x="Education (y
  theme_minimal() +
  geom_smooth(method="lm", se=FALSE, aes(color="OLS Line")) +
  scale_color_manual(name="", values="blue")
```

2.1 Derivation of OLS Estimators

How do we choose the “best” values for $\hat{\beta}_0$ and $\hat{\beta}_1$? We want a line that fits the data as closely as possible.

We define the **residual**, e_i, as the difference between the actual value y_i and the fitted value $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

The **Ordinary Least Squares (OLS)** method chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the **Sum of Squared Residuals (SSR)**:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We square the residuals so that positive and negative errors don’t cancel out, and because it penalizes larger errors more heavily.

Derivation of OLS (cont.)

To minimize the SSR, we use calculus: take the partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set them to zero. These are the **First Order Conditions (FOCs)**.

1. $\frac{\partial SSR}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \implies \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
2. $\frac{\partial SSR}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \implies \sum x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

Solving this system of two equations for the two unknowns $(\hat{\beta}_0, \hat{\beta}_1)$ gives the OLS estimator formulas:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x, y)}{\text{Sample Variance}(x)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y.

2.2 Algebraic Properties of OLS

The OLS estimators have some important algebraic properties that come directly from the FOCs:

1. **The sum of the OLS residuals is zero:**

$$\sum_{i=1}^{n} e_i = 0$$

   This implies that the sample average of the residuals, $\bar{e}$, is also zero.

2. **The sample covariance between the regressor (x) and the OLS residuals (e) is zero:**

$$\sum_{i=1}^{n} x_i e_i = 0$$

   This means the part of y that we can't explain with x (the residual) is uncorrelated with x in our sample.

3. **The point $(\bar{x}, \bar{y})$ is always on the OLS regression line.** From the formula

2.3 Interpreting OLS Coefficients

Let's run our wage-education regression: $\texttt{wage} = \hat{\beta}_0 + \hat{\beta}_1 \; \texttt{educ}$

```r
slr_model <- lm(wage ~ educ, data = dat)
# The coefficients are:
coef(slr_model)
```

```
## (Intercept)        educ
##    1.786342    1.149829
```

So our estimated SRF is: $\widehat{wage} = 1.79 + 1.15 \times educ$

Interpretation:

- Slope ($\hat{\beta}_1 \approx 1.25$): "For each additional year of education, we estimate the hourly wage to increase by **\$1.25**, on average." This is the key policy parameter.
- Intercept ($\hat{\beta}_0 \approx 0.84$): "For an individual with zero years of education, we predict an hourly wage of **\$0.84**."

2.4 Units and Functional Form

The values of the coefficients depend on the units of measurement of y and x. We've used a **level-level** model (y and x are in their natural units).

Suppose we measured wage in cents instead of dollars. * The new dependent variable is $wage_{cents} = 100 \times wage$. * The new regression would be: $\widehat{wage_{cents}} = (100 \times \hat{\beta}_0) + (100 \times \hat{\beta}_1) \times educ$ * Both the intercept and slope would be 100 times larger. The *interpretation* is the same, just the units change ("an extra year of education increases wage by 125 cents").

What if we measured education in months instead of years?

- The interpretation of $\hat{\beta}_1$ would become "the estimated change in wage for an additional *month* of education." The coefficient value would be $\frac{1}{12}$ of its original value.

2.5 Goodness-of-Fit

How well does our estimated line explain the variation in our dependent variable, y?

We can partition the total variation in y into two parts: the part explained by the model, and the part that is not explained.

- **SST (Total Sum of Squares):** Total variation in y. $SST = \sum(y_i - \bar{y})^2$
- **SSE (Explained Sum of Squares):** Variation explained by the regression. $SSE = \sum(\hat{y}_i - \bar{y})^2$
- **SSR (Sum of Squared Residuals):** Unexplained variation. $SSR = \sum e_i^2$

It is a mathematical property that **SST = SSE + SSR**.

Goodness-of-Fit: R-squared and SER

R-squared (R^2)

The **R-squared** measures the proportion of the total sample variation in y that is “explained” by the regression model.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- R^2 is always between 0 and 1.
- A higher R^2 means the model fits the data better in-sample.
- **Caution:** A high R^2 is not the ultimate goal of econometrics! We care more about getting an unbiased estimate of the causal effect β_1.

Standard Error of the Regression (SER)

The **SER** is an estimator of the standard deviation of the population error term, σ. It measures the typical size of a residual (the model's "average mistake").

$$\hat{\sigma} = SER = \sqrt{\frac{SSR}{n-2}}$$

* We divide by $n-2$ (degrees of freedom) because we had to estimate two parameters (β_0, β_1) to get the residuals. * SER is measured in the same units as y. A smaller SER is better.

2.6 Understanding Statistical Output

Let's look at the full output from R/Python/Stata for our simple regression.

```
summary(slr_model)
```

```
## 
## Call:
## lm(formula = wage ~ educ, data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.7201 -2.3878 -0.3926  1.9554 11.6092 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   1.7863     2.5164   0.710    0.479    
## educ          1.1498     0.1887   6.092 2.19e-08 ***
```

Understanding Statistical Output (Cont.)

Coefficients:

- `Estimate`: These are $\hat{\beta}_0$ (Intercept) and $\hat{\beta}_1$ (educ).
- `Std. Error`: The standard errors of the estimates, $se(\hat{\beta}_j)$, which measure their sampling uncertainty.
- `t value`: The t-statistic used for hypothesis testing (`Estimate` / `Std. Error`).
- `Pr(>|t|)`: The p-value for the t-test.

Goodness-of-Fit:

- `Residual standard error`: This is the **SER** (3.4).
- `Multiple R-squared`: This is our **R-squared** (0.27). The model explains about 46% of the variation in wages.

3. The Classical Assumptions (for SLR)

For our OLS estimates to have desirable statistical properties, certain assumptions must hold. These are the **Gauss-Markov Assumptions**.

- Assumption 1: Linearity in Parameters. The population model is $y = \beta_0 + \beta_1 x + u$.
- Assumption 2: Random Sampling. The data (x_i, y_i) are a random sample from the population described by the model.
- Assumption 3: Sample Variation in x. The values of x_i in the sample are not all the same. This is the **no perfect collinearity** assumption. If all x_i are the same, the denominator of $\hat{\beta}_1$ is zero!
- Assumption 4: Zero Conditional Mean. $E(u|x) = 0$. The average value of the unobserved factors is unrelated to the value of x.
- Assumption 5: Homoskedasticity. $Var(u|x) = \sigma^2$. The variance of the error term is constant for all values of x.

3.2 The Crucial Assumption: Zero Conditional Mean

$$E(u|x) = 0$$

This is the most important assumption for establishing **causality**. It means that the explanatory variable (x) must not be correlated with any of the unobserved factors (u) that affect the dependent variable (y).

Example: `wage` on `educ`. * $y = wage$ * $x = educ$ * $u =$ unobserved factors like innate ability, family background, motivation...

Is it likely that $E(u|educ) = 0$? * Probably not. Innate ability (u) is likely correlated with education (x). People with higher ability may find it easier to get more education. * If $Cov(educ, ability) > 0$, then our OLS estimate $\hat{\beta}_1$ will be biased upwards. It will capture the effect of education *and* the effect of ability. This is **Omitted Variable Bias**.

Unbiasedness of OLS

Theorem: Under assumptions **SLR.1 through SLR.4**, the OLS estimators are **unbiased**.

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad \mathbb{E}(\hat{\beta}_1) = \beta_1$$

What does this mean? * Unbiasedness is a property of the *procedure* of OLS estimation. * If we could draw many, many random samples from the population and calculate $\hat{\beta}_1$ for each sample, the *average* of all these estimates would be equal to the true population parameter, β_1. * Our estimate from any single sample might be higher or lower than the true value, but on average, we get it right. * This property relies critically on the Zero Conditional Mean assumption (SLR.4). If SLR.4 fails, OLS is biased.

Variance of OLS Estimators

We also want our estimators to be precise, meaning they don't vary too much from sample to sample. This is measured by their sampling variance.

Theorem: Under assumptions **SLR.1 through SLR.5** (all five Gauss-Markov assumptions), the variance of the OLS slope estimator is:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

What determines the precision of our estimate?

1. **The error variance,** σ^2: More "noise" in the relationship (larger σ^2) leads to a larger variance for $\hat{\beta}_1$.
2. **The total sample variation in x,** SST_x: More variation in our explanatory variable (x) leads to a *smaller* variance for $\hat{\beta}_1$. We learn more about the slope when our x values are more spread out.
3. **The sample size,** n: A larger sample size generally increases SST_x, which

4. Introduction to Multiple Linear Regression

Simple Linear Regression is often inadequate because we can't control for other factors that might be important. This leads to omitted variable bias.

The solution is to include those other factors in the model. This is **Multiple Linear Regression (MLR)**.

The Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

Now we have k explanatory variables.

Interpretation of Coefficients: - β_j is the effect of a one-unit change in x_j on y, **holding all other explanatory variables ($x_1, ..., x_{j-1}, x_{j+1}, ...x_k$) constant.** - This is the concept of *ceteris paribus* (all else equal). MLR allows us to isolate the effect of one variable while mathematically controlling for the others.

4.1 OLS Estimation in MLR

The principle is the same: we choose $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ to minimize the Sum of Squared Residuals (SSR). The formulas are complex (usually done with matrix algebra) but are easily handled by software.

Variance of OLS Estimators in MLR

The variance of a coefficient $\hat{\beta}_j$ now depends on **multicollinearity** – how correlated x_j is with the *other* explanatory variables.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

- SST_j is the total variation in x_j. - R_j^2 is the R-squared from a regression of x_j on all other explanatory variables in the model. - If x_j is highly correlated with other x's, R_j^2 will be close to 1, making the denominator small and $Var(\hat{\beta}_j)$ very large. This is **imperfect multicollinearity**. - The **no perfect collinearity** assumption for MLR means that no x_j can be a perfect linear combination of the others (i.e., $R_j^2 \neq 1$).

Heuristic Derivation of the t-statistic

To test a hypothesis about a single coefficient (e.g., $H_0 : \beta_j = 0$), we want to see how many standard deviations our estimate $\hat{\beta}_j$ is from the hypothesized value.

A standardized statistic would look like this:

$$\text{Statistic} = \frac{\text{Estimate} - \text{Hypothesized Value}}{\text{Standard Deviation of Estimate}} = \frac{\hat{\beta}_j - 0}{sd(\hat{\beta}_j)}$$

If we knew the population standard deviation $sd(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{SST_j(1-R_j^2)}}$, this statistic would follow a standard normal (Z) distribution.

The Problem: We don't know the population error variance, σ^2.

The Solution: We replace σ^2 with its sample estimate, $\hat{\sigma}^2 = \frac{SSR}{n-k-1}$. * This gives us the **standard error** of the estimate, $se(\hat{\beta}_j)$. * $se(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}} = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}}$

Why a t-statistic? (From Z to t)

Because we had to *estimate* σ^2, we introduce extra sampling variability into our statistic.

The ratio of our estimate to its standard error is no longer normally distributed. It follows a **t-distribution**.

$$t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

- The **t-distribution** looks very similar to the normal distribution but has "fatter tails," reflecting the added uncertainty from estimating σ^2.
- It is characterized by its **degrees of freedom (df)**, which for MLR is $df = n - k - 1$.
  - $n =$ sample size
  - $k + 1 =$ number of estimated parameters (including the intercept)
- As the sample size (n) gets large, the t-distribution converges to the standard normal distribution.

MLR Example: Wage, Education, and Experience

Let's add exper (years of work experience) to our model.

```
mlr_model <- lm(wage ~ educ + exper, data=dat_mlr)
summary(mlr_model)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = dat_mlr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3544 -1.9212  0.1218  2.1522  7.5290
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.82941    2.64978   1.068   0.2883
```