Ordinary least squares regression

Steven Kerr

Supported by

Ordinary least squares

- Let's say we have data consisting of n observations, y_i, x_i, $i = 1 \ldots n$.

- Example:
  - y_i = height of individual i.
  - x_i = weight of individual i.

- We wish to model the relationship between height and weight.

- Ordinary least squares is just about the simplest modelling strategy there is.

Ordinary least squares

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- β_0, β_1 are unknown parameters that we wish to estimate.
- ϵ_i is sometimes called the ‘error’.
- Let $\hat{\beta}_0, \hat{\beta}_1$ be possible values for β_0, β_1. Then we can estimate values for y_i too,
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
- The difference between the estimated values and the predicted values is called the *residual*,
$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$
- Roughly speaking, we wish to minimise the residuals across all observations.

Ordinary least squares

- Let's say we try to minimise the sum of the residuals,

$$\Sigma_{\mathrm{i}}(y_i - \hat{y}_i) = (y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + \dots (y_n - \hat{y}_n).$$

- The problem is that we could have some predicted values that are too large, and others too small, and they 'cancel each other out'.

- For example, let's say
  - Observation 1: $\mathrm{y}_1 = 1.85m$, $\hat{\mathrm{y}}_1 = 1.80m$.
  - Observation 2: $\mathrm{y}_2 = 1.70m$, $\hat{\mathrm{y}}_2 = 1.75m$.
  - $(y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) = 5 - 5 = 0.$

Ordinary least squares

- Instead, we choose $\hat{\beta}_0, \hat{\beta}_1$ to minimise the sum of the *square* of the residuals
$$\Sigma_i(y_i - \hat{y}_i)^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots (y_n - \hat{y}_n)^2.$$

- Hence the name *least squares*.

- We don't need to limit ourselves to just one predictor, or linear functions of the predictors.

- E.g. Let $z_i = 1$ if individual i is female, and 0 if they are male. We could fit the model
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 z_i + \epsilon_i,$$

  or any other model we think is sensible.

The dot product

- We often use a more compact notation.
- Let $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$ and $x_i = \begin{pmatrix} x_{i0} \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}$. These are called *vectors*.
- We can write the model in terms of the *dot product*,

$$y_i = \beta \cdot x_i + \epsilon_i = x_{i0}\beta_0 + x_{i1}\beta_1 + \ \ldots x_{ik}\beta_k + \epsilon_i.$$

Ordinary least squares

- It is possible to show that under some reasonable assumptions (beyond the scope of this course), when the sample is large the OLS estimator $\hat{\beta}$ *converges in probability* to β, and $\hat{\beta}$ *is* approximately normally distributed (central limit theorem).

- This means, roughly speaking, that as the dataset that we use to make our estimates gets bigger, our estimates get closer and closer to the 'true' values.

- Many models have theoretical results like this.

- In sparklyr, use the following command to fit a model using OLS:

  ml_linear_regression(data, formula)