Estimation for Regression Models

STAT 340: Applied Regression

Linear (least squares) regression

Ordinary least squares

Intuition: best model has small residual sum of squares (RSS)

Choose $\hat{\boldsymbol{\beta}}$ to minimize RSS:

$$min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Leftrightarrow min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2.$$

If we want to minimize this, first we need to find a critical point:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 = 2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))(-1) \\
0 &= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 = 2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))(-x_{i1}) \\
&\vdots \\
0 &= \frac{\partial}{\partial \beta_p} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 = 2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))(-x_{ip})
\end{aligned}$$

Now, let's move things around to solve for $\boldsymbol{\beta}$:

$$\begin{aligned}
\Rightarrow \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})(1) &= \sum_{i=1}^{n} y_i(1) \\
\sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})(x_{i1}) &= \sum_{i=1}^{n} y_i(x_{i1}) \\
&\cdots \\
\sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})(x_{ip}) &= \sum_{i=1}^{n} y_i(x_{ip})
\end{aligned}$$

Let's reorganize to get this in matrix form:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$
$$\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}'\mathbf{y}$$

$$\begin{aligned} &\Rightarrow (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &\Rightarrow \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

Are we done? No, we would have to verify that this is a minimum by showing the Hessian is positive definite (the second derivative is positive).

So, our estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Maximum likelihood

Rather than thinking about minimizing the RSS, as above, equivalently, we can maximize the likelihood to find $\hat{\boldsymbol{\beta}}$. The likelihood for a normal linear model, with mean $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ and variance σ^2 is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2}.$$

As above, we use calculus to find the maximum (this time). Maximizing the likelihood is the same thing as maximizing the log likelihood:

$$max_{\boldsymbol{\beta}} \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \right\} = max_{\boldsymbol{\beta}} \left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \right\}$$

You can verify that the solution is the same as shown above using ordinary least squares by setting the derivatives equal to 0 and checking the second derivative.

The advantage of maximum likelihood over ordinary least squares is that it is more versatile - we can use it whenever we have a likelihood that we can write down (which will be true for everything we do in this class). It will be really helpful for generalized linear models.

Examples

Example 1

Suppose we fit a linear model with no explanatory variables - only an intercept:

For $i = 1, ..., n$,

$$y_i = \beta + \epsilon_i,$$
$$\epsilon_i \sim Normal(0, \sigma^2)$$

.

(a) Write down the design (model) matrix, X.

(b) Find $\hat{\beta}$.

(c) Verify the result for (b) using R.

Example 2

Suppose we fit a one-way ANOVA model:

For $i = 1, ..., n$,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
$$x_i = \begin{cases} 0 & \text{if obs. i in first treatment group} \\ 1 & \text{if obs. i in second treatment group} \end{cases}$$

Suppose we have $n = 3$ observations, with observation 1 in the first treatment group and observations 2 and 3 in the second treatment group.

(a) Write down the design matrix X.

(b) Find $\hat{\beta}$.

(c) Verify the result for (b) using R.