Elements of statistical learning: Chapter 3

July 29, 2020

Content

1. Linear models
   - Sampling properties of $\hat{\beta}$
   - Gauss-Markov Theorem

2. Multiple regression
   - From simple univariate to multiple regressions

LS estimator

Let $\boldsymbol{X}$ be an $N \times (p+1)$ matrix of explanatory variables and $\boldsymbol{y}$ an $N \times 1$ vector of outputs. Then we know the LS estimator $\hat{\beta}$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

[see lecture slides "ESL1" for recap and proof].

The "hat" matrix

As such for the fitted linear model

$$\begin{aligned} \hat{y} &= \boldsymbol{X}\hat{\beta} \\ &= \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{\boldsymbol{H}}\boldsymbol{y} \\ &= \boldsymbol{H}\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{H}$ is commonly referred to as the hat matrix.

$\boldsymbol{H}$ the projection matrix

Let us denote the column vectors of $\boldsymbol{X}$ by $\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ with $\boldsymbol{x}_0 \equiv 1$.

- These vectors span a subspace of $\mathbb{R}^N$, also referred to as a column vector of $\boldsymbol{X}$.
- We minimize $RSS(\beta) = ||\boldsymbol{y} - \boldsymbol{X}\beta||^2$ by choosing $\hat{\beta}$, so that the residual vector $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to this subspace.
- the hat matrix $\boldsymbol{H}$ computes the orthogonal projection, and hence it is also known as the projection matrix.

Variance-covariance matrix

Assumptions

1. Observations y_i are uncorrelated have constant variance σ^2
2. x_i are fixed (i.e. non-stochastic)

$$
\begin{aligned}
var(\hat{\beta}) &= var\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\
&= var\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta+\epsilon)\right] \\
&= var\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\beta+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon\right] \\
&= var\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon\right] \\
&= \mathbb{E}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon]'\right\} \\
&= \mathbb{E}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon\epsilon'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\} \\
&= \mathbb{E}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\epsilon\epsilon'(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}
\end{aligned}
$$

Note that ϵ is the error term and has zero mean and also remember that $\boldsymbol{X}$ is fixed, and thus

$$\mathbb{E}[aZ] = a\mathbb{E}[Z]$$

where Z is a random variable and a is a constant. Therefore,

$$\begin{aligned} var(\hat{\beta}) &= \mathbb{E}\left\{\epsilon\epsilon'(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}E\left\{\epsilon\epsilon'\right\} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2 \end{aligned}$$

where σ^2 can be calculated by

$$\sigma^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

thus, assuming the errors are further Gaussian

$$\hat{\beta} \sim N(\beta, (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2)$$

Gauss-Markov Theorem

Least squares estimator of parameter β has the smallest variance among all linear unbiased estimators. Why is the LS estimator unbiased?

Proof.

$$
\begin{aligned}
\hat{\beta} &= \mathbb{E}[\hat{\beta}] \\
&= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \\
&= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{\epsilon})] \\
&= \mathbb{E}[\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}] \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{\epsilon}] \\
&= \beta
\end{aligned}
$$

□

From simple univariate to muliple regressions

Suppose first we have a univariate model with no intercept

$$Y_i = X_i \beta + \varepsilon_i$$

The least squares estimates and residuals are

$$\hat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

with residuals

$$r_i = y_i - x_i \hat{\beta}$$

which in vector notation can be expressed as

$$< \boldsymbol{x}, \boldsymbol{y} > = \sum_{i=1}^{N} x_i y_i = \boldsymbol{x}' \boldsymbol{y}$$

which is the inner product between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Thus, the OLS estimator $\hat{\beta}$ can be expressed as follows

$$\hat{\beta} = \frac{< \boldsymbol{x}, \boldsymbol{y} >}{< \boldsymbol{x}, \boldsymbol{x} >},$$

Suppose now that we have p inputs $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$, which are the columns of the matrix $\boldsymbol{X}$ and are orthogonal, such that $< \boldsymbol{x}_j, \boldsymbol{x}_k >= 0$ for all $i \neq j$. When the inputs are orthogonal, the multiple least squares estimates $\hat{\beta}_j$ are equal tothe univariate estimates - i.e.

$$\hat{\beta}_j = \frac{< \boldsymbol{x}_j, \boldsymbol{y} >}{< \boldsymbol{x}_j, \boldsymbol{x}_j >}$$

In other words, the inputs are orthogonal and have no impact on each other's parameters estimates in the model.

Consider the case of an intercept and a single input $\boldsymbol{x}$, then the least squares coefficient of $\boldsymbol{x}$ has the form

$$\hat{\beta}_1 = \frac{< \boldsymbol{x} - \bar{x}1, \boldsymbol{y} >}{< \boldsymbol{x} - \bar{x}1, \boldsymbol{x} - \bar{x}1 >}$$

The steps of the algorithm can be seen as follows

1. Regress $\boldsymbol{x}$ on 1 to obtain $\bar{x}1$
2. Obtain the residuals $\boldsymbol{z} = \boldsymbol{x} - \bar{x}1$
3. Regress $\boldsymbol{y}$ on $\boldsymbol{z}$ to obtain the coefficient $\hat{\beta}_1$
$$\hat{\beta}_1 = \frac{< \boldsymbol{z}, \boldsymbol{y} >}{< \boldsymbol{z}, \boldsymbol{z} >}$$

Step 1 orthogonalizes x with respect to $x_o = 1$.