STAT 425

Multiple Linear Regression. Part 2

Properties of the Least-Square estimates

- In MLR the LS estimate $\hat{\boldsymbol{\beta}}$ is a random vector, since it is a function of $\mathbf{y}$
- For hypothesis testing we want to find the probability distribution of $\hat{\boldsymbol{\beta}}$
- We will review the definitions of the mean and variance of a random vector first, and how to calculate mean and variances of affine transformations of this random vector.

Review: Mean and Variances of Random Vectors

Mean of a Random Vector

Let $\mathbf{Z}$ a random vector of size $m \times 1$, with components $Z_1, Z_2, \dots, Z_m$. The mean of $\mathbf{Z}$ is equal to vector $\boldsymbol{\mu}$ defined as:

$$\boldsymbol{\mu} = E[\mathbf{Z}] = \begin{pmatrix} E[Z_1] \\ E[Z_2] \\ \dots \\ E[Z_m] \end{pmatrix}$$

Variance of a Random Vector

The Variance of a random vector $\mathbf{Z}$ is a matrix (the Variance-Covariance matrix). This matrix is symmetric of size $m \times m$ with component (i, j) equal to the $Cov(Z_i, Z_j)$

$$\begin{aligned}\Sigma_{m\times m} = Cov(\mathbf{Z}) &= E[(\mathbf{Z} - \boldsymbol{\mu})(\mathbf{Z} - \boldsymbol{\mu})^t] \\ &= \begin{pmatrix} Var(Z_1) & \dots & Cov(Z_1, Z_m) \\ \dots & \dots & \dots \\ Cov(Z_m, Z_1) & \dots & Var(Z_m) \end{pmatrix}\end{aligned}$$

Mean and Covariance matrix of an affine transformation

Assume an affine transformation of the form:

$$\mathbf{W} = \mathbf{a_{n\times 1}} + \mathbf{B_{n\times m}Z_{m\times 1}}$$

$$E[\mathbf{W}] = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \;\; Cov(\mathbf{W}) = \mathbf{B\Sigma B}^t$$

In particular consider a transformation of the form:

$$W = \mathbf{v^t Z} = v_1 Z_1 + v_2 Z_2 + \ldots + v_m Z_m$$

$$E[W] = \mathbf{v}^t\boldsymbol{\mu} = \sum_{i=1}^{m} v_i \mu_i$$

$$Var(W) = \mathbf{v}^t \Sigma \mathbf{v} = \sum_{i=1}^{m} v_i^2 Var(Z_i) + 2\sum_{i<j} v_i v_j Cov(Z_i, Z_j)$$

Mean and Covariance of the LS estimates

We use the assumption: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with

$$E[\mathbf{e}] = \mathbf{0}, \;\; Cov(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

Using these assumptions we get:

$$E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}, \;\; Cov(\mathbf{y}) = \sigma^2 \mathbf{I}_n$$

Under these assumptions we also get:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\mathbf{X^t X})^{-\mathbf{1}}\mathbf{X^t y}] \\ &= (\mathbf{X^t X})^{-\mathbf{1}}\mathbf{X^t} E[\mathbf{y}] \\ &= (\mathbf{X^t X})^{-\mathbf{1}}\mathbf{X^t X}\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned}$$

We have shown that the LS estimate is unbiased.

Mean and Covariance of the LS estimates (Cont.)

$$\begin{aligned}
\mathsf{Cov}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathsf{Cov}(\mathbf{y})\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right]^t \\
&= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\sigma^2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1};
\end{aligned}$$

Using the previous results we can also show:

$$E[\hat{\mathbf{y}}] = \mathbf{X}\boldsymbol{\beta}, \;\; Cov(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$$
$$E[\mathbf{r}] = \mathbf{0}, \;\; Cov(\mathbf{r}) = \sigma^2(\mathbf{I_n} - \mathbf{H})$$
$$E[\hat{\sigma}^2] = \frac{1}{n-p}E[\mathbf{r^t r}] = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2$$

[1]

[1] It can be shown that $\frac{\mathbf{r^t r}}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$

- $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are unbiased estimators of $\boldsymbol{\beta}$ and σ^2 respectively
- We can plug-in the variance estimator $\hat{\sigma}^2$ to get the covariance of $\hat{\boldsymbol{\beta}}$
- The standard errors of the $\hat{\beta}_i$ are the square roots of the elements of the diagonal of the covariance matrix $Cov(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-1}$. For example:

$$se(\hat{\beta}_1) = \hat{\sigma}\sqrt{[(\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-1}]_{11}}$$

The Gauss-Markov Theorem

The main reason why we use LS estimation is because of the Gauss-Markov theorem. If the errors are uncorrelated, have equal variance and mean equal to zero, the LS estimators have a very nice property: they have the lowest variance within the class of linear estimators.

- Suppose we are interested in estimating a linear combination of $\boldsymbol{\beta}$ of the form:

  $$\theta = \mathbf{c}^t \boldsymbol{\beta} = \sum_{j=1}^{p} c_j \beta_j$$

  For example, estimating any element of $\boldsymbol{\beta}$ and estimating the mean response at a new value x^* are all special cases of this setup.

The Gauss-Markov Theorem (Cont.)

- Naturally, we can form an estimate of θ by plugging in the LS estimate $\boldsymbol{\beta}$ in the equation for θ:

  $$\hat{\theta}_{LS} = \mathbf{c}^t\hat{\boldsymbol{\beta}} = \mathbf{c^t}(\mathbf{X^tX})^{-\mathbf{1}}\mathbf{X^ty}$$

  This is a linear[2] and unbiased estimator of θ. Its mean square error can be calculated as:

  $$MSE(\hat{\theta}_{LS}) = E[\hat{\theta}_{LS} - \theta]^2 = Var(\hat{\theta}_{LS})$$

[2] It is a linear combination of the n data points $y_1, y_2, \ldots, y_n$

Gauss-Markov theorem (Cont.)

- Suppose there is another estimate of θ, which is also linear and unbiased. The following Theorem states that $\hat{\theta}_{LS}$ is always better in the sense that its MSE is always smaller (or at least, not bigger)
- Gauss-Markov Theorem: The estimator $\hat{\theta}_{LS} = \mathbf{c}^t\hat{\boldsymbol{\beta}}$ is the BLUE (best linear unbiased estimator) of the parameter $\mathbf{c}^t\boldsymbol{\beta}$ for any vector $\mathbf{c} \in \Re^p$.

Proof: Please see Supplemental Material.

Maximum Likelihood Estimation

Recall the normal assumptions for the regression model:

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + e_i \ (i = 1, \ldots, n)$$

with $e_i \sim N(0, \sigma^2)$. This implies:

$$\mathbf{y} \sim \mathbf{N_n}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

We can show that the likelihood function can be written as:

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto \frac{RSS}{n}^{-\frac{n}{2}}$$

The value of $\boldsymbol{\beta}$ that maximizes the Likelihood function is the Maximum Likelihood Estimator (MLE) of $\boldsymbol{\beta}$. This estimator is equal to the LS estimate of $\boldsymbol{\beta}$.

Distributions of the Least-Squares estimates

Recall the assumption for the linear regression model:

$$\mathbf{y} \sim \mathbf{N_n}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

Any affine transformation of $\mathbf{y}$ will also have a Normal distribution[3]. We can use the identities to calculate the mean and variance of an affine transformation of a random vector to get the following results:

$$\begin{aligned}\hat{\boldsymbol{\beta}} = (\mathbf{X^tX})^{-1}\mathbf{X^ty} &\sim \mathbf{N_p}(\boldsymbol{\beta}, \sigma^2(\mathbf{X^tX})^{-1}) \\ \hat{\mathbf{y}} = \mathbf{Hy} &\sim \mathbf{N_n}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}) \\ \mathbf{r} = (\mathbf{I_n} - \mathbf{H})\mathbf{y} &\sim \mathbf{N_n}(\mathbf{0}, \sigma^2(\mathbf{I_n} - \mathbf{H}))\end{aligned}$$

[3] They will also have a joint Normal distribution

Note that for the fitted values $\hat{\mathbf{y}}$ and the estimated residuals $\hat{\mathrm{e}} = \mathbf{r}$ we can calculate the mean and covariance matrices as follows:

$$\begin{aligned}
E[\hat{\mathbf{y}}] &= \mathbf{H}E[\mathbf{y}] = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} \\
Cov(\hat{\mathbf{y}}) &= \mathbf{H}\sigma^2\mathbf{H}^t = \sigma^2\mathbf{H} \\
E[\mathbf{r}] &= (\mathbf{I_n} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \\
Cov(\mathbf{r}) &= (\mathbf{I_n} - \mathbf{H})\sigma^2(\mathbf{I_n} - \mathbf{H})^t = \sigma^2(\mathbf{I_n} - \mathbf{H})
\end{aligned}$$

- Although $\mathbf{r}$ is a vector of dimension n, it always lies in a subspace of dimension $(n-p)$.
- $\mathbf{r}$ behaves like a random vector with a distribution $\mathbf{N}_{n-p}(\mathbf{0}, \sigma^2\mathbf{I}_{n-p})$, so we have:

$$\hat{\sigma}^2 = \frac{||\mathbf{r}||^2}{n-p} \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}$$

- It can be show that $\hat{\mathbf{y}}$ and $\mathbf{r}$ are uncorrelated since they are in orthogonal spaces. Since they have a joint normal distribution, they are independent.[4]

[4] Note that if two random variables are uncorrelated, they are not necessarily independent, unless they have a joint Normal distribution