Relevant statistics for the method of Least Squares

We assume that the observed values y_i can be written on the following form

$$y_i = \widehat{y}_i + \epsilon_i$$

where $\widehat{y}_i$ is the prediction and ϵ_i is the error. We assume that the ϵ_i's are random variables that are independently and identically distributed from a normal distribution $N\left(\mu = 0, \sigma^2\right)$. This makes the observed values y_i independent random variables as well. The mean of y_i is

$$\langle y_i \rangle = \widehat{y}_i$$

and the variance of y_i is

$$\mathrm{Var}(y_i) = \sigma^2$$

See Excercise 3.pdf. Since the parameters $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ depend on $\mathbf{y}$ whose components are random variables, the components of $\widehat{\boldsymbol{\beta}}$ are random variables also. We denote the expectation value of $\widehat{\boldsymbol{\beta}}$ as $\boldsymbol{\beta}$. The covariance matrix of the components of $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \mathbb{E}[\widehat{\boldsymbol{\beta}}])(\widehat{\boldsymbol{\beta}} - \mathbb{E}[\widehat{\boldsymbol{\beta}}])^T\right]$$

$$= \mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right] = \mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\left(\widehat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}^T\right)\right]$$

$$= \mathbb{E}\left[\widehat{\boldsymbol{\beta}\boldsymbol{\beta}}^T - \widehat{\boldsymbol{\beta}}\boldsymbol{\beta}^T - \boldsymbol{\beta}\widehat{\boldsymbol{\beta}}^T + \boldsymbol{\beta}\boldsymbol{\beta}^T\right] = \mathbb{E}\left[\widehat{\boldsymbol{\beta}\boldsymbol{\beta}}^T\right] - \mathbb{E}[\widehat{\boldsymbol{\beta}}]\boldsymbol{\beta}^T - \boldsymbol{\beta}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}^T\right] + \boldsymbol{\beta}\boldsymbol{\beta}^T]$$

$$= \mathbb{E}\left[\widehat{\boldsymbol{\beta}\boldsymbol{\beta}}^T\right] - \boldsymbol{\beta}\boldsymbol{\beta}^T - \boldsymbol{\beta}\boldsymbol{\beta}^T + \boldsymbol{\beta}\boldsymbol{\beta}^T = \mathbb{E}\left[\widehat{\boldsymbol{\beta}\boldsymbol{\beta}}^T\right] - \boldsymbol{\beta}\boldsymbol{\beta}^T$$

$$= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right\}^T\right] - \boldsymbol{\beta}\boldsymbol{\beta}^T$$

$$= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right] - \boldsymbol{\beta}\boldsymbol{\beta}^T$$

In the last line it has been used that $(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$ and that

$\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right\}^T = \left\{\left(\mathbf{X}^T\mathbf{X}\right)^T\right\}^{-1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ (the transpose of the inverse of a matrix is equal to the inverse of the transpose of a matrix). Now we have

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right]\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^T$$

Where the expectation value $\mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right]$ is

$$\mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] = \mathbb{E}\left[(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})^T\right]$$

$$= \mathbb{E}\left[(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})\left(\boldsymbol{\beta}^T\mathbf{X}^T+\boldsymbol{\epsilon}^T\right)\right] = \mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}\boldsymbol{\beta}^T\mathbf{X}^T + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]$$

$$= \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \mathbf{X}\boldsymbol{\beta}\mathbb{E}\left[\boldsymbol{\epsilon}^T\right] + \mathbb{E}[\boldsymbol{\epsilon}]\boldsymbol{\beta}^T\mathbf{X}^T + \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] = \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]$$

The last term is the covariance matrix of $\boldsymbol{\epsilon}$ since

$$\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] = \mathbb{E}\left[(\boldsymbol{\epsilon}-\mathbf{0})(\boldsymbol{\epsilon}-\mathbf{0})^T\right] = \mathbb{E}\left[(\boldsymbol{\epsilon}-\mathbb{E}[\boldsymbol{\epsilon}])(\boldsymbol{\epsilon}-\mathbb{E}[\boldsymbol{\epsilon}])^T\right] = \mathrm{Var}(\boldsymbol{\epsilon})$$

The diagonal elements of $\mathrm{Var}(\boldsymbol{\epsilon})$ are

$$\mathrm{Var}(\boldsymbol{\epsilon})_{ii} = \mathbb{E}\left[(\epsilon_i - \mathbb{E}[\epsilon_i])^2\right] = \mathrm{Var}(\epsilon_i) = \sigma^2$$

The non-diagonal elements of $\mathrm{Var}(\boldsymbol{\epsilon})$ are

$$\mathrm{Var}(\boldsymbol{\epsilon})_{ij} = \mathbb{E}[(\epsilon_i - \mathbb{E}[\epsilon_i])(\epsilon_j - \mathbb{E}[\epsilon_j])]$$

$$= \mathbb{E}[(\epsilon_i - 0)(\epsilon_j - 0)] = \mathbb{E}[\epsilon_i\epsilon_j]$$

$$= \mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j] = 0$$

where the last line follows from the assumption that each ϵ_i are independent (but identically distributed) variables. Therefore $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ and

$$\mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] = \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \sigma^2\mathbf{I}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \sigma^2\mathbf{I}\right)\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^T$$

$$= \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^{T}\mathbf{X}^{T}\mathbf{X}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1} + \sigma^{2}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^{T}$$

$$= \boldsymbol{\beta}\boldsymbol{\beta}^{T} + \sigma^{2}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{\mathbf{-1}} - \boldsymbol{\beta}\boldsymbol{\beta}^{T} = \sigma^{2}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{\mathbf{-1}}$$

In particular the variance of the components of $\widehat{\boldsymbol{\beta}}$ are the diagonal elements of the covariance matrix

$$\mathrm{Var}(\widehat{\beta}_{i}) = \sigma^{2}\left[\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\right]_{ii}$$