Chapter 3. Multiple Linear Regression - Proofs

Preliminary on Multivariate Normal Distribution

A p-dimensional random vector $\mathbf{X} = (X_1, \cdots, X_p)'$ follows a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a mean vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)'$ and a covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{1 \le i,j \le p}$ means that *each entry* of $\mathbf{X}$ has a marginal normal distribution

$$X_j \sim N(\mu_j, \sigma_{jj}) \quad \text{for all } j = 1, \cdots, p$$

Meanwhile, the covariance between the entries of $\mathbf{X}$ is given by $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$ for every pair of $\mathbf{X}$ with $1 \le i \ne j \le p$.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \right)$$

Lemma

If the p-dimensional normal random vector $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A} \in \mathbb{R}^{q \times p}$, then the linear transformation $\mathbf{AX}$ is also q-dimensional normal random vector with distribution $\mathbf{AX} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

Chapter 3. Multiple Linear Regression - Proofs

Least Squares Estimate $\hat{\beta}$

The gradient of the sum of squares $S(\boldsymbol{\beta})$ is $\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = 2\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} - \mathbf{y})$. Set the gradient to zero at the solution $\hat{\boldsymbol{\beta}}$, we get

$$\mathbf{X}'(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}) = \mathbf{0} \quad \Longrightarrow \quad (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Multiplying both sides by the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$, we obtain

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{is a } (p+1)\text{-dim vector}$$

is the least squares estimate.

Chapter 3. Multiple Linear Regression - Proofs

Statistical Properties of $\hat{\beta}$

Denote the n-dimensional vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)'$ containing the random error terms from all observations. Then, the multilinear regression model gives that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

implying that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$.

The randomness comes from $\boldsymbol{\varepsilon} \sim N(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$

Here $\mathbf{0}_n$ is an all zero n-dimensional vector and $\mathbf{I}_n$ is an $n \times n$ identity matrix.

Therefore, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$ is a normal random vector following distribution $N(\mathbf{0}_{p+1}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. This suggests that the least squares **estimator** $\hat{\boldsymbol{\beta}}$ follows a multivariate normal distribution:

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

This implies that the j-th entry of $\hat{\boldsymbol{\beta}}$ also follows a normal distribution

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj})$$

where $[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}$ is the (j, j)-th entry of $(\mathbf{X}'\mathbf{X})^{-1}$.

Chapter 3. Multiple Linear Regression - Proofs

Unbiasedness of $\hat{\sigma}^2$

By definition, the fitted value

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

then the residuals

$$\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} y_1 - \hat{y}_i \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is an $n \times n$ matrix and $\mathbf{H}$ is a projection matrix meaning that $\mathbf{H}^2 = \mathbf{H}$.
Therefore, by definition and the fact $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we get

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\|\mathbf{e}\|^2}{n-p-1} = \frac{\mathbf{e}'\mathbf{e}}{n-p-1} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p-1} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p-1} \\ &= \underbrace{\frac{\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta}}{n-p-1}}_{=0} + \underbrace{\frac{2\boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p-1}}_{\text{its expectation=0}} + \frac{\boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}}{n-p-1}\end{aligned}$$

Therefore, we have

$$E(\hat{\sigma}^2) = \frac{E[\boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}]}{n-p-1} = \frac{E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}] - E[\boldsymbol{\varepsilon}'\mathbf{H}\boldsymbol{\varepsilon}]}{n-p-1} = \frac{\sigma^2(n-p-1)}{n-p-1} = \sigma^2$$

Chapter 3. Multiple Linear Regression - Proofs

SST = SSR + SSE

By definition, denote $\bar{\mathbf{y}} = (\bar{y}, \cdots, \bar{y})'$, we have

$$\begin{aligned}\mathrm{SST} =& \|\mathbf{y} - \bar{\mathbf{y}}\|^2 = \|(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \bar{\mathbf{y}})\|^2 \\ =& \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 + 2\big\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} - \bar{\mathbf{y}}\big\rangle \\ =& \mathrm{SSE} + \mathrm{SSR} + 2\big\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} - \bar{\mathbf{y}}\big\rangle\end{aligned}$$

It suffices to show that $\big\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} - \bar{\mathbf{y}}\big\rangle = 0$.

Recall from the previous slide that $\mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, implying that $\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}}\rangle = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{y} = \mathbf{y}'\mathbf{0}\mathbf{y} = 0$ where we used the fact $\mathbf{H} = \mathbf{H}^2$. In addition, $\langle \mathbf{y} - \hat{\mathbf{y}}, \bar{\mathbf{y}}\rangle = \bar{y} \cdot \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{1}_n\rangle = \bar{y} \cdot \sum_{i=1}^n (y_i - \hat{y}_i) = 0$ (why?).

Essentially, $\hat{\mathbf{y}} - \bar{\mathbf{y}} = \mathbf{H}(\mathbf{y} - \bar{\mathbf{y}})$ is the projection of $\mathbf{y} - \bar{\mathbf{y}}$ onto the column space of $\mathbf{X}$, and $\mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})(\mathbf{y} - \bar{\mathbf{y}})$ is the projection of $\mathbf{y} - \bar{\mathbf{y}}$ onto the complement of the column space of $\mathbf{X}$, i.e., which can not be explained by the columns of $\mathbf{X}$.