5 Inference under the linear model

Recall that the LS estimator is given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}, \quad \boldsymbol{A} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top.$$

The corresponding vectors of *fitted (predicted) values* and *residuals* are given, respectively, by

$$\hat{\mathbf{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \qquad \boldsymbol{e} = \mathbf{Y} - \hat{\mathbf{Y}},$$

and we have $\boldsymbol{e} \perp \hat{\mathbf{Y}}$ (this was a *defining* property of the LS solution). Remember that all of this holds regardless of the assumptions of the linear model, and in fact requires no statistical assumptions at all.

Now, assume the linear model (16). Then the vector $\mathbf{Y}$ becomes a *random* vector, with distribution generally depending on the unknown *parameters* $\boldsymbol{\beta}$ and σ^2. Same goes for $\hat{\boldsymbol{\beta}}$, which now has a meaning as an *esimator* of the unknown parameter $\boldsymbol{\beta}$, the true (unknown) coefficient vector. Let us calculate its mean and covariance matrix. We have

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{\beta}] + \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{X}\boldsymbol{\beta}$$

and

$$\operatorname{cov}(\boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Also, we can now calculate

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{A}\mathbb{E}\boldsymbol{Y} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\beta},$$

and

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \operatorname{cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\operatorname{cov}(\boldsymbol{Y})\boldsymbol{A}^\top = \boldsymbol{A}\left(\sigma^2 \boldsymbol{I}\right)\boldsymbol{A}^\top &= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right]^\top \\ &= \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}.\end{aligned}$$

Hence

$$\boxed{\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \qquad \operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}. \qquad (16)}$$

Estimating the noise level σ^2. Intuitively, it makes sense to use the residual vector $\boldsymbol{e}$ to estimate σ^2. Define

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\|\boldsymbol{e}\|^2 = \frac{1}{n-p-1}\sum_{i=1}^n e_i^2.$$

Proposition 8. *$\hat{\sigma}^2$ defined above is an unbiased estimator of σ^2.*

Proof. Let $M = \operatorname{Im}(\boldsymbol{X})$. Denote $\boldsymbol{Q} := \boldsymbol{I} - \boldsymbol{P}$ for the projection matrix onto $M^\perp$. Then

$$\boldsymbol{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \boldsymbol{P}\mathbf{Y} = (\boldsymbol{I} - \boldsymbol{P})\mathbf{Y} = \boldsymbol{Q}\boldsymbol{Y},$$

and note that we also have

$$\boldsymbol{Q}\mathbf{Y} = \boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\epsilon} = \boldsymbol{Q}\boldsymbol{\epsilon},$$

because $\boldsymbol{QX} = \mathbf{0}$ (every column of $\boldsymbol{X}$ is in $\operatorname{Im}(\boldsymbol{X})$), so we can conclude $\boldsymbol{e} = \boldsymbol{Q\epsilon}$. Therefore,

$$\|\boldsymbol{e}\|^2 = \|\boldsymbol{Q\epsilon}\|^2 = \boldsymbol{\epsilon}^\top \boldsymbol{Q}^\top \boldsymbol{Q\epsilon} = \boldsymbol{\epsilon}^\top \boldsymbol{Q\epsilon} = \sum_i \sum_j Q_{ij}\epsilon_i\epsilon_j,$$

since $\boldsymbol{Q}$ is symmetric and idempotent (projection matrix), and so

$$\mathbb{E}\|\boldsymbol{e}\|^2 = \mathbb{E}\|\boldsymbol{Q\epsilon}\|^2 = \mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}^\top \boldsymbol{Q\epsilon}\right] = \mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q\epsilon}\right] = \mathbb{E}\left[\sum_i \sum_j Q_{ij}\epsilon_i\epsilon_j\right] = \sum_i \sum_j Q_{ij}\mathbb{E}\left[\epsilon_i\epsilon_j\right]. \quad (17)$$

Also, because $\mathbb{E}\epsilon_i = 0$ by assumption, we have

$$\mathbb{E}\left[\epsilon_i\epsilon_j\right] = \operatorname{Cov}\left(\epsilon_i, \epsilon_j\right) = \left\{ \begin{array}{ll} \sigma^2, & i = j \\ 0, & \text{otherwise} \end{array} \right. .$$

Continuing from (17),

$$\mathbb{E}\|\boldsymbol{e}\|^2 = \sum_i \sum_j Q_{ij}\mathbb{E}\left[\epsilon_i\epsilon_j\right] = \sum_i Q_{ii}\mathbb{E}\left[\epsilon_i^2\right] = \sum_i Q_{ii}V\left(\epsilon_i\right) = \sigma^2 \sum_i Q_{ii} = \sigma^2 \operatorname{tr}(\boldsymbol{Q}). \quad (18)$$

Now, we known that, as a projection matrix, $\boldsymbol{Q}$ is similar to a diagonal matrix $\boldsymbol{D}$ whose diagonal has $\dim(\boldsymbol{Q}) = n - (p+1) = n - p - 1$ entries equal to 1 , and the rest $p + 1$ entries are zero. But the trace is preserved under the similarity relation, meaning that $\operatorname{tr}(Q) = \operatorname{tr}(\boldsymbol{D}) = n - p - 1$. Continuing from (18), we get

$$\mathbb{E}\|\boldsymbol{e}\|^2 = \sigma^2 \operatorname{tr}(\boldsymbol{Q}) = \sigma^2(n - p - 1)$$

implying

$$\mathbb{E}\left[\frac{1}{n-p-1}\|\boldsymbol{e}\|^2\right] = \sigma^2.$$

□

We can give an alternative, shorter proof using the following general result.

Lemma 1. *For any random vector $\boldsymbol{Z}$ it holds that*

$$\mathbb{E}\|\boldsymbol{Z}\|^2 = \operatorname{tr}\left(\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right]\right) = \operatorname{tr}\left(\operatorname{cov}(\boldsymbol{Z}) + \boldsymbol{\mu_Z}\boldsymbol{\mu_Z}^\top\right),$$

where $\boldsymbol{\mu_Z} := \mathbb{E}\boldsymbol{Z}$. As a special case, if $\mu_{\boldsymbol{Z}} = \mathbf{0}$, then $\mathbb{E}\|\boldsymbol{Z}\|^2 = \operatorname{tr}(\operatorname{cov}(\boldsymbol{Z}))$.

Proof of lemma. . We have

$$\mathbb{E}\|\boldsymbol{Z}\|^2 = \mathbb{E}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right) \overset{(a)}{=} \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)\right] \overset{(b)}{=} \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)\right] \overset{(c)}{=} \operatorname{tr}\left(\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right]\right) \overset{(d)}{=} \operatorname{tr}\left(\operatorname{cov}(\boldsymbol{Z}) + \boldsymbol{\mu_Z}\boldsymbol{\mu_Z}^\top\right)$$

where (a) is because $\boldsymbol{Z}^\top\boldsymbol{Z}$ is a scalar; (b) is due to the general identity $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$; (c) is due to the definition of the expectation of a random matrix, and the linearity of the expectation; and (d) is due to the general identity $\operatorname{cov}(\boldsymbol{Z}) = \mathbb{E}\left[\mathbf{ZZ}^\top\right] - \boldsymbol{\mu}_Z\boldsymbol{\mu}_{\mathbf{Z}}^\top$. □

Alternative proof of Proposition 8. . Taking $\boldsymbol{Z} = \boldsymbol{e}$ in the lemma above, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}\|^2 = \operatorname{tr}\left(\operatorname{cov}(\boldsymbol{e}) + \boldsymbol{\mu}_{\boldsymbol{e}}\boldsymbol{\mu}_{\boldsymbol{e}}^\top\right) \overset{(a)}{=} \operatorname{tr}(\operatorname{cov}(\boldsymbol{e})) = \\ \operatorname{tr}(\operatorname{cov}(\boldsymbol{Q}\boldsymbol{Y})) = \operatorname{tr}\left(\boldsymbol{Q}\operatorname{cov}(\boldsymbol{Y})\boldsymbol{Q}^\top\right) = \operatorname{tr}\left(\boldsymbol{Q}\left[\sigma^2\boldsymbol{I}\right]\boldsymbol{Q}^\top\right) = \\ \sigma^2\operatorname{tr}\left(\boldsymbol{Q}\boldsymbol{Q}^\top\right) = \sigma^2\operatorname{tr}(\boldsymbol{Q}) = \sigma^2(n-p-1)\end{aligned}$$

where (a) is because $\mathbb{E}\boldsymbol{e} = \mathbb{E}[\boldsymbol{Q}\mathbf{Y}] = \boldsymbol{Q}\mathbb{E}\mathbf{Y} = \boldsymbol{Q}\mathbb{E}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{Q}[\mathbb{E}\boldsymbol{\epsilon}] = \mathbf{0}$, and the last steps are as in the original proof we gave. □