Linear combinations. If $\boldsymbol{a} = (a_0, \ldots, a_p)^\top \in \mathbb{R}^{p+1}$ is a fixed vector, then

$$\theta := \boldsymbol{a}^\top \boldsymbol{\beta} = \sum_{j=0}^{p} a_j \beta_j \in \mathbb{R} \tag{19}$$

is called a *linear combination* (of $\boldsymbol{\beta}$). Consider estimating a linear combination (19). A natural estimator is

$$\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}} = \boldsymbol{a}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{c}^\top \boldsymbol{Y}, \tag{20}$$

where

$$\boldsymbol{c} := \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{a}, \tag{21}$$

and note that $\boldsymbol{c} \in \mathbb{R}^n$ whereas $\boldsymbol{a} \in \mathbb{R}^{p+1}$. This estimator is a linear function of $\mathbf{Y}$, and we can calculate its mean and variance under the linear model,

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}] = \boldsymbol{a}^\top \mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{a}^\top \boldsymbol{\beta} = \theta$$

and

$$V(\hat{\theta}) = V(\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}) = V\left(\boldsymbol{c}^\top \boldsymbol{Y}\right) = \boldsymbol{c}^\top \operatorname{cov}(\boldsymbol{Y})\boldsymbol{c} = \boldsymbol{c}^\top \left[\sigma^2 \mathbf{I}_n\right] \boldsymbol{c} = \sigma^2 \boldsymbol{c}^\top \boldsymbol{c}.$$

Thus $\hat{\theta}$ is a *linear, unbiased* estimator of θ with variance $\sigma^2 \boldsymbol{c}^\top \boldsymbol{c}$. Is there are *better* linear unbiased estimator of θ? First we need to define "better". The mean squared error (MSE) of an estimator $\hat{\theta}$ of θ is

$$\operatorname{MSE}(\hat{\theta}) := \mathbb{E}_\theta[(\hat{\theta} - \theta)^2],$$

and notice that this generally depends on the true value θ. We will say that an estimator $\hat{\theta}$ of θ is *better* than another estimator $\tilde{\theta}$ if

$$\operatorname{MSE}(\hat{\theta}) \leq \operatorname{MSE}(\tilde{\theta}) \qquad \text{for all } \theta.$$

For any estimator $\hat{\theta}$, we have

$$\begin{aligned}
\operatorname{MSE}(\hat{\theta}) &= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)^2\right] = \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right] + \mathbb{E}\left[(\mathbb{E}\hat{\theta} - \theta)^2\right] + 2\underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)]}_{=0} = \\
&= \underbrace{\mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right]}_{V(\hat{\theta})} + \underbrace{\mathbb{E}\left[(\mathbb{E}\hat{\theta} - \theta)^2\right]}_{(\operatorname{bias}(\hat{\theta}))^2}
\end{aligned}$$

where we used that fact that $\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta}) = \mathbb{E}\hat{\theta} - \mathbb{E}\hat{\theta} = 0$.

We conclude from the general decomposition (5) that an unbiased estimator has

$$\operatorname{MSE}(\hat{\theta}) = V(\hat{\theta}).$$

Hence, restricting attention to unbiased estimators, the estimator $\hat{\theta}$ is better than another estimator $\tilde{\theta}$ if

$$V(\hat{\theta}) \leq V(\tilde{\theta}) \quad \forall \theta.$$

The following theorem, maybe the most famous result in all of linear regression, says that, under the linear model (15), the LS estimator $\hat{\boldsymbol{\beta}}$ is the *best linear unbiased estimator* (BLUE) of θ.

Theorem 1 (Gauss-Markov)**.** . *Let $\theta := \boldsymbol{a}^\top \boldsymbol{\beta}$ be a linear combination, and assume the linear model* (15)*. Denote by $\hat{\theta}$ the LS estimator of θ in* (20)*, and consider another linear unbiased estimator $\hat{\theta}$ of θ*

$$\tilde{\theta} = \boldsymbol{d}^\top \boldsymbol{Y}, \quad \mathbb{E}[\tilde{\theta}] = \theta \quad \forall \theta.$$

Then

$$V(\hat{\theta}) \leq V(\tilde{\theta}) \quad \forall \theta$$

Proof. For c defined in (21), write

$$\boldsymbol{d} = \boldsymbol{c} + \boldsymbol{\Delta}, \quad \boldsymbol{\Delta} := \boldsymbol{d} - \boldsymbol{c} \in \mathbb{R}^n.$$

$\tilde{\theta}$ is unbiased, hence for all $\boldsymbol{\beta}$ we have

$$\begin{aligned} \theta = \mathbb{E}\tilde{\theta} = \mathbb{E}\left[\boldsymbol{d}^\top \boldsymbol{Y}\right] = \mathbb{E}\left[(\boldsymbol{c} + \boldsymbol{\Delta})^\top \boldsymbol{Y}\right] = \mathbb{E}\left[(\boldsymbol{c} + \boldsymbol{\Delta})^\top \boldsymbol{Y}\right] &= \mathbb{E}\left[\left(\boldsymbol{c}^\top \boldsymbol{Y}\right] + \mathbb{E}\left[\left(\boldsymbol{\Delta}^\top \boldsymbol{Y}\right]\right.\right. \\ &= \theta + \boldsymbol{\Delta}^\top \mathbb{E}[\boldsymbol{Y}] = \theta + \boldsymbol{\Delta}^\top \boldsymbol{X}\boldsymbol{\beta}, \end{aligned}$$

where the second-to-last equality is due to unbiasedness of $\hat{\theta}$. Comparing the two extreme sides of the sequence of the equality, we get

$$\boldsymbol{\Delta}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \quad \forall \boldsymbol{\beta} \quad \Rightarrow \quad \boldsymbol{\Delta}^\top \boldsymbol{X} = \boldsymbol{0},$$

so

$$\boldsymbol{\Delta}^\top \boldsymbol{c} = \underbrace{\boldsymbol{\Delta}^\top \boldsymbol{X}}_{=0} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{a} = \boldsymbol{0}.$$

We then calculate

$$\begin{aligned} V(\tilde{\theta}) = V\left(\boldsymbol{d}^\top \boldsymbol{Y}\right) &= V\left[(\boldsymbol{c} + \boldsymbol{\Delta})^\top \mathbf{Y}\right] \\ &= \operatorname{cov}\left[(\boldsymbol{c} + \boldsymbol{\Delta})^\top \mathbf{Y}\right] \\ &= (\boldsymbol{c} + \boldsymbol{\Delta})^\top \operatorname{cov}[\boldsymbol{Y}](\boldsymbol{c} + \boldsymbol{\Delta}) \\ &= (\boldsymbol{c} + \boldsymbol{\Delta})^\top \sigma^2 \left[\boldsymbol{I}_n\right](\boldsymbol{c} + \boldsymbol{\Delta}) \\ &= \sigma^2 (\boldsymbol{c} + \boldsymbol{\Delta})^\top (\boldsymbol{c} + \boldsymbol{\Delta}) \\ &= \sigma^2 \left(\boldsymbol{c}^\top \boldsymbol{c} + \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\right) \\ &\geq \sigma^2 \boldsymbol{c}^\top \boldsymbol{c} \\ &= V(\hat{\theta}). \end{aligned}$$

□

We have considered point estimation of a scalar $\theta = \theta(\boldsymbol{\beta})$, more specifically unbiased estimation of a linear function of $\boldsymbol{\beta}$. We now want to move on to other inferential tasks, for example we'll want to use the LS estimator $\hat{\boldsymbol{\beta}}$ to construct a confidence interval for $\boldsymbol{\beta}$, or to test whether a particular coordinate β_j is equal to zero. For this we will need some further assumptions on the linear model.

Review of multivariate distributions. All the concepts presented here generalize naturally beyond the two dimensional case. If Z_1, Z_2 are two random variables, then $Z = (Z_1, Z_2)^\top$ is a random vector of dimension 2. The joint cumulative distribution function (CDF) of $\mathbf{Z}$ is

$$F_{\boldsymbol{Z}}(z_1, z_2) := P(Z_1 \le z_1, Z_2 \le z_2),$$

which is always defined and determines the distribution of $\boldsymbol{Z}$. The variables Z_1 and Z_2 are (statistically) independent if

$$F_{\boldsymbol{Z}}(z_1, z_2) = P(Z_1 \le z_1) P(Z_2 \le z_2) \qquad \text{for all } z_1, z_2 \in \mathbb{R}.$$

If the derivative

$$f_{\mathbf{Z}}(z_1, z_2) = \frac{\partial^2}{\partial z_1 z_2} F_{\mathbf{Z}}(z_1, z_2)$$

exists (for all except maybe a subset of $\mathbb{R}^2$ of probability zero), we call $f_{\mathbf{Z}}$ the *joint density* of $\boldsymbol{Z}$, and we have the relation

$$F_{\mathbf{Z}}(z_1, z_2) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} f_{\mathbf{Z}}(u_1, u_2)\, du_1 du_2.$$

Of course, the derivative is in that case an equivalent characterization of the distribution of Z.

The multivariate Normal distribution.

Definition 4. We say that a random vector $\boldsymbol{W} = (W_1, \dots, W_k)^\top$ has a multivariate normal distribution if there exists a representation

$$\boldsymbol{W} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{A}\mathbf{Z} \tag{22}$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{A} \in \mathbb{R}^{k \times l}$ are constant (nonrandom) and where $\boldsymbol{Z} = (Z_1, \dots, Z_l)^\top$ is a random vector whose components Z_i are i.i.d. $\mathcal{N}(0, 1)$ random variables ("$\stackrel{d}{=}$" means "equal in distribution").

Properties of the multivariate Normal distribution.

1. If $\boldsymbol{W}$ has a multivariate normal distribution, then

$$\mathbb{E}[\boldsymbol{W}] = \mathbb{E}[\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{\mu}] + \mathbb{E}[\boldsymbol{A}\boldsymbol{Z}] = \boldsymbol{\mu} + \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu}$$
$$\mathrm{cov}(\boldsymbol{W}) = \mathrm{cov}(\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}) = \mathrm{cov}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\,\mathrm{cov}(\boldsymbol{Z})\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top$$

Therefore, if there is another representation $\boldsymbol{W} \stackrel{d}{=} \boldsymbol{\mu'} + \boldsymbol{A'}\mathbf{Z}$, then necessarily $\boldsymbol{\mu}' = \boldsymbol{\mu}$ and $\boldsymbol{A}'\boldsymbol{A}'^\top = \boldsymbol{A}\boldsymbol{A}^\top$ (this, in turn, can be shown to hold if and only if $\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{U}^\top$ for an orthogonal matrix $\boldsymbol{U}$ – try to prove this).

2. In (22) suppose that $l = k$, and if $\boldsymbol{A}_{k\times k}$ has linearly independent columns, and denote $\boldsymbol{V} := \boldsymbol{A}\boldsymbol{A}^\top$. Then

$$f_W(\boldsymbol{w}) = (2\pi)^{-m/2}|\boldsymbol{V}|^{-1/2} \exp\left[-(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{w} - \boldsymbol{\mu})/2\right], \quad \boldsymbol{w} \in \mathbb{R}^m$$

Hence, the distribution of $\boldsymbol{W}$ in (22) is completely determined by $\boldsymbol{\mu}$ and $\boldsymbol{V}$.

We write

$$\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$$

for the multivariate distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ (this notation applies whether or not $\boldsymbol{A}_{k\times k}$ has linearly independent columns).

3. It is a consequence of 2 that if $\boldsymbol{W}^{(1)} = \boldsymbol{\mu} + \boldsymbol{A}^{(1)}\mathbf{Z}^{(1)}$ and $\boldsymbol{W}^{(2)} = \boldsymbol{\mu} + \boldsymbol{A}^{(2)}\mathbf{Z}^{(2)}$, and if $\boldsymbol{A}^{(1)}\boldsymbol{A}^{(1)\top} = \boldsymbol{A}^{(2)}\boldsymbol{A}^{(2)\top}$, then $\boldsymbol{W}^{(1)} \stackrel{d}{=} \boldsymbol{W}^{(2)} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$.

4. From the previous properties, if $\boldsymbol{c} \in \mathbb{R}^k$ is a constant vector, then

$$\boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V} \boldsymbol{c}\right)$$

In words, a linear combination of a multivariate normal vector has a univariate normal distribution. In particular, if we take $c = (\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots, 0}_{k-j})^\top$, then

$$W_j = \boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\mu_j, \boldsymbol{V}_{jj}\right)$$

5. If for a random vector $\boldsymbol{W}$ it holds that $\boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V} \boldsymbol{c}\right) \forall \boldsymbol{c} \in \mathbb{R}^m$, where $\boldsymbol{\mu}$ and $\boldsymbol{V}$ denote the mean and covariance of $\boldsymbol{W}$, then $\boldsymbol{W}$ has a multivariate normal distribution. Combined with property 4 , this says

$$\boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V} \boldsymbol{c}\right) \quad \forall \boldsymbol{c} \in \mathbb{R}^m \quad \Longleftrightarrow \quad \boldsymbol{W} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{V}).$$

Thus, Property 4 is in fact a *defining property* of the multivariate normal distribution.

5. If $\boldsymbol{C} \in \mathbb{R}^{m \times k}$ constant matrix then $\boldsymbol{C}\boldsymbol{W} \sim \mathscr{N}_m\left(\boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{V}\boldsymbol{C}^T\right)$.

6. If $\boldsymbol{W}^{(j)} \sim \mathcal{N}_k\left(\boldsymbol{\mu}^{(j)}, \boldsymbol{V}^{(j)}\right), j = 1, \ldots, p$, independent, and if d_j are scalar constants, then

$$\sum_{j=1}^{p} d_j \boldsymbol{W}^{(j)} \sim \mathcal{N}_k\left(\sum_{j=1}^{p} d_j \boldsymbol{\mu}^{(j)}, \sum_{j=1}^{p} d_j^2 \boldsymbol{V}^{(j)}\right)$$

7. Let $\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$ and $\mathscr{I}_1, \mathscr{I}_2 \subseteq \{1, \ldots, k\}$ disjoint subsets of indices. If $\operatorname{Cov}(W_i, W_j) = 0 \quad \forall i \in \mathscr{I}_1, j \in \mathscr{I}_2$, then the vectors

$$\boldsymbol{W}^{(1)} = (W_l : l \in \mathscr{I}_2) \in \mathbb{R}^{|\mathcal{I}_2|}, \quad \boldsymbol{W}^{(2)} = (W_k : k \in \mathscr{I}_1) \in \mathbb{R}^{|\mathscr{I}_1|}$$

are statistically independent.

Distributions related to the normal.

Definition 5 (Chi-square distribution)**.** If $Z_1, Z_2, \ldots, Z_k \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$, then the distribution of

$$Q = \sum_{j=1}^{k} Z_j^2$$

is called the Chi-square distribution with k degrees of freedom, and we denote $Q \sim \chi_k^2$ (in R: `pchisq()`, `qchisg()`, `rchisq()`).

Definition 6 (t-distribution)**.** . If $Z \sim \mathcal{N}(0, 1), V \sim \chi_k^2$, are independent random variables, then the distribution of

$$T = \frac{Z}{\sqrt{V/k}}$$

is called the t-distribution with k degrees of freedom, and we denote $T \sim t_k$ (in R: `pt ()`, `qt( )`, `rt( )`).

Definition 7 (F distribution)**.** If $V_1 \sim \chi^2_{k_1}, V_2 \sim \chi^2_{k_2}$, are independent random variables, the distribution of

$$F = \frac{V_1/k_1}{V_2/k_2}$$

is called the F*-distribution with* k_1 *and* k_2 (numerator and denominator, respectively) *degrees of freedom*, and we denote $F \sim \mathrm{F}_{k_1,k_2}$.

Proposition. If $Q \sim \chi^2_k$, then $\mathbb{E}Q = k$.

Proof. For $Z_i \sim \mathcal{N}(0,1)$, iid for $i = 1, \ldots, k$, we can write $Q \overset{d}{=} \sum_{i=1}^k Z_i^2$, where " $\overset{d}{=}$ " means "equal in distribution". Then $\mathbb{E}Q \overset{d}{=} \mathbb{E}\sum_{i=1}^k Z_i^2 = \sum_{i=1}^k \mathbb{E}Z_i^2 = \sum_{i=1}^k V(Z_i) = k$. □

Proposition. Let $\boldsymbol{Z} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I})$, and $\boldsymbol{P}$ be a square symmetric $\left(\boldsymbol{P}^\top = \boldsymbol{P}\right)$ and idempotent $\left(\boldsymbol{P}^2 = \boldsymbol{P}\right)$ matrix with $\operatorname{rank}(\boldsymbol{P}) = r$. Then $\|\boldsymbol{P}Z\|^2 \sim \chi^2_r$.

Proof. From a previous lemma, since $\mathbb{E}[\boldsymbol{P}\boldsymbol{Z}] = \boldsymbol{P}\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$, we have $\mathbb{E}\|\boldsymbol{P}\boldsymbol{Z}\|^2 = \operatorname{tr}(\operatorname{cov}[\boldsymbol{P}\boldsymbol{Z}]) = \operatorname{tr}\left(\boldsymbol{P}\boldsymbol{I}\boldsymbol{P}^\top\right) = \operatorname{tr}(\boldsymbol{P}) = r$, where the last equality is because $\boldsymbol{P}$ is similar to a diagonal matrix with r nonzero elements on its diagonal. □

Inference under the normal linear model.

Recall:

The linear model: $\quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}, \quad \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$

We will now make the additional assumption that the error term $\boldsymbol{\epsilon}$ has a *multivariate normal* distribution. In other words, we will assume The normal linear model:

The *normal* linear model: $\quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathscr{N}_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_n\right)$

The additional normality assumption will enable us to address inferential tasks beyond point estimation, e.g., to construct a confidence interval for a linear combination of $\hat{\boldsymbol{\beta}}$. Indeed, if we assume $\boldsymbol{\epsilon}$ has a multivariate normal distribution, then we can derive exact distributions of $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$, and their joint.

Distribution of $\hat{\boldsymbol{\beta}}$. Recall that, for $\boldsymbol{A} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \in \mathbb{R}^{(p+1)\times n}$, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} &= \boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon} \\
&\overset{d}{=} \boldsymbol{\beta} + (\sigma\boldsymbol{A})\mathbf{Z}
\end{aligned}$$

where $\boldsymbol{Z} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I})$. Hence, by definition, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution. We have already calculated the moments of $\hat{\boldsymbol{\beta}}$,

$$\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \quad \operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1},$$

so in conclusion we have

$$\hat{\boldsymbol{\beta}} \sim \mathscr{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

Distribution of $\hat{\sigma}^2$. Recall that $\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{\epsilon}$, where $\boldsymbol{Q}$ is the $n \times n$ projection matrix onto the orthogonal complement of $\operatorname{Im}(\boldsymbol{X})$. By a previous result, $\|\boldsymbol{e}\|^2 \sim \sigma^2\chi^2_{n-p-1}$. This gives

$$\frac{n-p-1}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p-1} \quad \Longleftrightarrow \quad \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-p-1}}{n-p-1} \quad \sim t_{n-p-1}.$$

Joint distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. For $\boldsymbol{A} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \in \mathbb{R}^{(p+1)\times n}$, first note that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} = \boldsymbol{A}\left(\boldsymbol{P}_M\boldsymbol{Y} + \boldsymbol{P}_{M^\perp}\boldsymbol{Y}\right) = \boldsymbol{P}\boldsymbol{P}\boldsymbol{P}_M\boldsymbol{Y} + \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{M}^\perp}\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{M}}\boldsymbol{Y}$$

Then,

$$\begin{aligned} \operatorname{cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \operatorname{cov}\left(\boldsymbol{A}\boldsymbol{P}_M\boldsymbol{Y}, \left(\boldsymbol{I}_n - \boldsymbol{P}_M\right)\boldsymbol{Y}\right) &= \boldsymbol{A}\boldsymbol{P}_M \operatorname{cov}(\boldsymbol{Y})\left(\boldsymbol{I}_n - \boldsymbol{P}_M\right)^\top \\ &= \sigma^2\boldsymbol{A}\boldsymbol{P}_M\left(\boldsymbol{I}_n - \boldsymbol{P}_M\right) = \boldsymbol{0} \end{aligned} \tag{23}$$

Moreover,

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_n - \boldsymbol{P}_M \end{bmatrix} \boldsymbol{Y} \overset{d}{=} \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_n - \boldsymbol{P}_M \end{bmatrix} (\boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{Z}) \tag{24}$$

i.e., $\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{e} \end{bmatrix}$ has a multivariate normal distribution. Together, (23) and (24) imply that $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent (because uncorrelated=independent under joint normality).