Statistical inference for a single coefficient.

I. Confidence interval. By the derivations above, we have

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 V_{jj}}} \sim \mathcal{N}(0,1), \quad \boldsymbol{V} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

Since σ^2 is unknown, we replace it naturally by its estimator $\hat{\sigma}^2$:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} = \frac{\left(\hat{\beta}_j - \beta_j\right)/\sqrt{\sigma^2 V_{jj}}}{\sqrt{\hat{\sigma}^2 V_{jj}}/\sqrt{\sigma^2 V_{jj}}} = \frac{\left(\hat{\beta}_j - \beta_j\right)/\sqrt{\sigma^2 V_{jj}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-p-1}/(n-p-1)}}$$

By definition, then,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \sim t_{n-p-1}. \tag{25}$$

Hence,

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(t_{n-p-1;\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \leq t_{n-p-1;1-\alpha/2}\right) \\
&= \mathbb{P}\left(\hat{\beta}_j - \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;\alpha/2} \leq \beta_j \leq \hat{\beta}_j + \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;1-\alpha/2}\right)
\end{aligned}$$

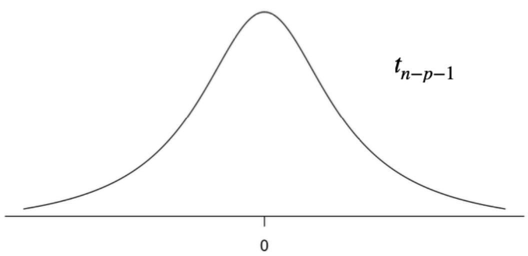

Figure 5: t distribution

In other words,

$$\boxed{CI = \left(\hat{\beta}_j - \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;\alpha/2}, \hat{\beta}_j + \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;1-\alpha/2}\right) \tag{26}}$$

is a confidence interval (CI) of level $1-\alpha$ for β_j.

More generally, let $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$ be a linear combination. Then, for the BLUE estimator $\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$, we have

$$\mathbb{E}[\hat{\theta}] = \boldsymbol{a}^\top \boldsymbol{\beta} = \theta, \qquad V(\hat{\theta}) = \operatorname{cov}\left(\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}\right) = \sigma^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}.$$

Therefore,

$$\frac{\hat{\theta}-\theta}{\sqrt{\sigma^2 a^\top V a}} \sim \mathcal{N}(0,1), \quad \boldsymbol{V} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \tag{27}$$

Repeating the argument from before,

$$\frac{\hat{\theta}-\theta}{\sqrt{\hat{\sigma}^2 a^\top V \boldsymbol{a}}} = \frac{(\hat{\theta}-\theta)/\sqrt{\sigma^2 \boldsymbol{a}^\top V \boldsymbol{a}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{(\hat{\theta}-\theta)/\sqrt{\sigma^2 a^\top V \boldsymbol{a}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}/(n-p-1)}},$$

so

$$\frac{\hat{\theta}-\theta}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}} \sim t_{n-p-1}. \tag{28}$$

In conclusion,

$$\boxed{CI = \hat{\theta} \pm t_{n-p-1;1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}} \tag{29}$$

is a Cl of level $1-\alpha$ for θ.

Note: for $\boldsymbol{a} = \boldsymbol{e}_j = (0,\ldots,0,1,0,\ldots,0)^\top$ we get $\theta = \beta_j$ and (29) coincides with (26).

II. Hypothesis testing. Based on (25), we can also derive a statistical test for

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

Indeed,

$$T_j := \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \overset{H_0}{\sim} t_{n-p-1} \Longrightarrow \mathbb{P}_{H_0}\left(t_{n-p-1;\alpha/2} \leq \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \leq t_{n-p-1;1-\alpha/2}\right) = 1-\alpha,$$

i.e., a level- α test is to reject H_0 if

$$|T_j| := |\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 V_{jj}}}| \geq t_{n-p-1;1-\alpha/2}. \tag{30}$$

Extending this to a general linear combination, let $\theta := \boldsymbol{a}^\top \boldsymbol{\beta}$, and suppose we want to test

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0.$$

Then, by (28), under the null we have

$$\frac{\hat{\theta}}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}} \overset{H_0}{\sim} t_{n-p-1}.$$

By a similar calculation to the one above, the test that rejects H_0 if

$$|\frac{\hat{\theta}}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}}| \geq t_{n-p-1;1-\alpha/2} \tag{31}$$

is a level-α test.

Note that any CI for θ naturally defines a test of

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq 0.$$

for *any* value θ_0. Indeed, let CI_α be a $(1-\alpha)$-level CI for θ, and consider the test that rejects H_0 whenever $\theta_0 \notin CI_\alpha$. Then, since a CI covers the true parameter θ for *all* θ, it holds in particular for $\theta = \theta_0$, and we have

$$P_{H_0}(\text{type I error}) = P_{\theta_0}(\theta_0 \notin CI_\alpha) \leq 1 - \alpha.$$

We emphasize that this holds for *any* valid CI for θ. If we choose the particular CI (29), then by definition the corresponding test rejects when

$$\theta_0 \notin \big(\hat{\theta} - t_{n-p-1;1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a}},\ \hat{\theta} + t_{n-p-1;1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a}}\big),$$

which is equivalent to rejecting H_0 if

$$|\frac{\hat{\theta} - \theta_0}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a}}}| \geq t_{n-p-1;1-\alpha/2}. \tag{32}$$

In the special case $\theta_0 = 0$, this coincides with (31). (Of course, since this holds for any linear combination θ, it holds in particular for $\theta = \beta_j$, and we get as a special case the duality between (26) and (30).

III. Prediction interval. Assume the normal linear model, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$. Now suppose we had a new independent observation $(\boldsymbol{X}_*, Y_*)$ from the same model, where $\boldsymbol{X}_* = (1, X_{*1}, ..., X_{*p})^\top \in \mathbb{R}^{p+1}$ is the feature (explanatory variables) vector and Y_* is the response for the new observation,

$$Y_* = \boldsymbol{X}_*^\top \boldsymbol{\beta} + \epsilon_*, \quad \epsilon_* \sim \mathcal{N}\left(0, \sigma^2\right).$$

We now want to construct a *prediction interval* for Y_*, i.e., a random interval $PI = PI\left(\boldsymbol{X}_*; \boldsymbol{X}, \boldsymbol{Y}\right)$ which is a function of the observed sample $(\boldsymbol{X}, \boldsymbol{Y})$, s.t.

$$\mathbb{P}\left(Y_* \in PI\right) = 1 - \alpha.$$

A natural *point predictor* for Y_* is

$$\hat{Y}_* = X_*^\top \hat{\boldsymbol{\beta}}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\hat{Y}_* - Y_*\right] = &\quad \mathbb{E}\hat{Y}_* - \mathbb{E}Y_* = \boldsymbol{X}_*^\top \mathbb{E}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_*^\top \boldsymbol{\beta} = 0 \\
V\left[\hat{Y}_* - Y_*\right] = &\quad V(\hat{Y}_*) + V(Y_*) - 2\operatorname{Cov}(\hat{Y}_*, Y_*) = \sigma^2 \boldsymbol{X}_*^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_* + \sigma^2 - 2\operatorname{Cov}(\boldsymbol{X}_*^\top \hat{\boldsymbol{\beta}}, \boldsymbol{X}_*^\top \boldsymbol{\beta} + \epsilon_*) = \\
&= \sigma^2 [1 + \boldsymbol{X}_*^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_*]
\end{aligned}$$

Moreover,

$$\begin{pmatrix} Y^* \\ \hat{Y}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} + \epsilon^* \\ \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} + \epsilon^* \\ \boldsymbol{x}^{*T}\boldsymbol{A}\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} + \epsilon^* \\ \boldsymbol{x}^{*T}\boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} \\ \boldsymbol{x}^{*T}\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} 1 & \boldsymbol{0} \\ 0 & x^{*T}\boldsymbol{A} \end{pmatrix} \begin{pmatrix} \epsilon^* \\ \boldsymbol{\epsilon} \end{pmatrix}$$

and, since ϵ^* and $\boldsymbol{\epsilon}$ are *independent* normals, we conclude that $\binom{Y^*}{\hat{Y}^*}$ has a multivariate (2-dim) normal distribution. This implies

$$Y^* - \hat{Y}^* \sim \mathcal{N}\big(0, \sigma^2 [1 + \boldsymbol{X}_*^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_*]\big),$$

because $Y^* - \hat{Y}^*$ is a linear transformation of $(\hat{Y}_*, Y_*)^\top$. With the usual argument replacing σ^2 with $\hat{\sigma}^2$, a $(1-\alpha)$-level prediction inverval for Y^* is then given by

$$\boxed{PI = \hat{Y}_* \pm t_{n-p-1;1-\alpha/2} \cdot \hat{\sigma}\sqrt{1 + \boldsymbol{X}_*^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_*}. \qquad (33)}$$

Least squares regression: demonstration with `R`. (in separate file)