Multicollinearity

Consider the usual regression equation

$$Y = X\beta + \epsilon$$

with $E(\epsilon \mid X) = \mathbf{0}$ and $\mathrm{Cov}(\epsilon \mid X) = \sigma^2 I$. Suppose that one column of X, say Z, is close to the span of the other columns of X. How much does this affect the SE of the corresponding coefficient estimator? Write $X\beta = W\alpha + Z\gamma$, where W is the matrix formed from the other columns of X besides Z. Then

$$Y = X\widehat{\beta} + e = W\widehat{\alpha} + Z\widehat{\gamma} + e \qquad \text{with} \qquad e \perp X\,. \tag{1}$$

We want to know how $\mathrm{SE}(\widehat{\gamma} \mid X)$ is affected by Z being close to $\mathrm{col}(W)$.

Suppose first that $Z \perp W$, which is the *opposite* of Z being close to $\mathrm{col}(W)$, unless $\|Z\|$ is small. We have $\mathrm{Cov}(\widehat{\beta} \mid X) = \sigma^2 (X'X)^{-1}$ and, supposing that Z is the last column of X,

$$X'X = [W\ Z]'[W\ Z] = \begin{bmatrix} W'W & W'Z \\ Z'W & Z'Z \end{bmatrix} = \begin{bmatrix} W'W & \mathbf{0} \\ \mathbf{0} & \|Z\|^2 \end{bmatrix}.$$

Thus,

$$(X'X)^{-1} = \begin{bmatrix} (W'W)^{-1} & \mathbf{0} \\ \mathbf{0} & 1/\|Z\|^2 \end{bmatrix}.$$

Therefore,

$$\mathrm{SE}(\widehat{\gamma} \mid X) = \frac{\sigma}{\|Z\|}\,. \tag{2}$$

In general, without assuming that $Z \perp W$, write

$$Z = P_W Z + P_W^{\perp} Z = Wb + P_W^{\perp} Z$$

for some b, where P_W denotes orthogonal projection onto $\mathrm{col}(W)$ and $P_W^{\perp}$ denotes orthogonal projection onto the orthocomplement of $\mathrm{col}(W)$. Using this, we may rewrite (1) as

$$Y = W\widehat{\alpha} + Z\widehat{\gamma} + e = W\widehat{\alpha} + Wb\widehat{\gamma} + P_W^{\perp} Z\widehat{\gamma} + e = W(\widehat{\alpha} + b\widehat{\gamma}) + P_W^{\perp} Z\widehat{\gamma} + e\,.$$

That is,

$$Y = W(\widehat{\alpha} + b\widehat{\gamma}) + P_W^{\perp} Z\widehat{\gamma} + e\,. \tag{3}$$

Since $e \perp W, Z$, we also have $e \perp P_W^{\perp} Z$:

$$0 = e \cdot Z = e \cdot (P_W Z + P_W^{\perp} Z) = e \cdot P_W^{\perp} Z\,.$$

Therefore, (3) is a regression of Y on W and $P_W^\perp Z$. In this regression, $\widehat{\gamma}$ is the coefficient of $P_W^\perp Z$. But by design, this regression now has $P_W^\perp Z \perp W$, whence our earlier formula (2) applies:

$$\mathrm{SE}(\widehat{\gamma} \mid X) = \frac{\sigma}{\|P_W^\perp Z\|} \,. \tag{4}$$

This is our answer: it shows that if closeness of Z to W is measured by $\|P_W^\perp Z\|$, then we get a precise measure of how such closeness affects $\mathrm{SE}(\widehat{\gamma} \mid X)$.

A formula that gives another interpretation is as follows. Define $R_{Z,W}^2$ to be the explained variance of regressing Z on W: we ignore whether there is an intercept or not and define it as

$$R_{Z,W}^2 := \frac{\|P_W Z\|^2}{\|Z\|^2} \,.$$

Since $\|P_W Z\|^2 + \|P_W^\perp Z\|^2 = \|Z\|^2$ by the Pythagorean theorem, we have $1 - R_{Z,W}^2 = \|P_W^\perp Z\|^2/\|Z\|^2$, whence (4) becomes

$$\mathrm{SE}(\widehat{\gamma} \mid X) = \frac{\sigma}{\sqrt{1 - R_{Z,W}^2} \cdot \|Z\|} \,.$$

In this formula, we can think of $\|Z\|$ as fixed and letting vary only the angle between Z and W, which amounts to varying $R_{Z,W}^2$. As the angle goes from $90°$ to $0°$, the explained variance $R_{Z,W}^2$ goes from 0 to 1. In fact, if θ is the angle between Z and W, then $R_{Z,W} = \cos\theta$ and $\sqrt{1 - R_{Z,W}^2} = \sin\theta$, as a picture shows.