For a general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ we have the general result that for the least-squares estimate

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

in the full rank case or, more generally,

$$\mathrm{var}(\mathbf{a}'\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a}$$

when $\mathbf{a}'\boldsymbol{\beta}$ is estimable. Therefore, how well we estimate our parameters depends on the error variance and the values of the predictors $\mathbf{X}$. In particular it does not depend on the data $\mathbf{Y}$(!).

ORTHOGONAL STRUCTURE IN THE DESIGN MATRIX

Partition the linear model as

$$E[\mathbf{Y}] = (\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_k)\begin{pmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{\beta}_1\\ \vdots\\ \boldsymbol{\beta}_k\end{pmatrix},$$

where $\mathbf{X}_j$ is $n \times p_j$, $\boldsymbol{\beta}_j$ is $p_j \times 1$, and $\sum_j p_j = p$. Suppose that the columns of $\mathbf{X}_i$ are orthogonal to those of $\mathbf{X}_j$, i.e.,

$$\mathbf{X}_i'\mathbf{X}_j = \mathbf{0}, \text{ for all } i, j.$$

Then $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ has the form

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}_0\\ \hat{\boldsymbol{\beta}}_1\\ \vdots\\ \hat{\boldsymbol{\beta}}_k\end{pmatrix} = \begin{pmatrix}(\mathbf{X}_0'\mathbf{X}_0)^{-1} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & (\mathbf{X}_k'\mathbf{X}_k)^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{X}_0'\mathbf{Y}\\ \mathbf{X}_1'\mathbf{Y}\\ \vdots\\ \mathbf{X}_k'\mathbf{Y}\end{pmatrix}$$

$$= \begin{pmatrix}(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Y}\\ (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}\\ \vdots\\ (\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}\end{pmatrix}.$$

Therefore, the least-squares estimate of $\boldsymbol{\beta}_i$ does not depend on any of the other terms are in the model.

Also,

$$\begin{aligned}
RSS &= (\mathbf{Y}-\hat{\mathbf{Y}})'(\mathbf{Y}-\hat{\mathbf{Y}}) \\
&= \mathbf{Y}'\mathbf{Y}-\hat{\mathbf{Y}}'\mathbf{Y}-\underbrace{\mathbf{Y}'}_{\mathbf{Y}=\hat{\mathbf{Y}}+\hat{\boldsymbol{\varepsilon}}}\hat{\mathbf{Y}}+\hat{\mathbf{Y}}'\hat{\mathbf{Y}} \\
&= \mathbf{Y}'\mathbf{Y}-\hat{\mathbf{Y}}'\mathbf{Y}-(\hat{\mathbf{Y}}+\hat{\boldsymbol{\varepsilon}})'\hat{\mathbf{Y}}+\hat{\mathbf{Y}}'\hat{\mathbf{Y}} \\
&= \mathbf{Y}'\mathbf{Y}-\hat{\mathbf{Y}}'\mathbf{Y}-\hat{\boldsymbol{\varepsilon}}'\hat{\mathbf{Y}} \\
&= \mathbf{Y}'\mathbf{Y}-\hat{\mathbf{Y}}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y}-\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y}-\sum_{i=0}^{k}\hat{\boldsymbol{\beta}}_i'\mathbf{X}_i'\mathbf{Y}
\end{aligned}$$

Therefore if $\boldsymbol{\beta}_i$ is set equal to 0, RSS increases by $\hat{\boldsymbol{\beta}}_i'\mathbf{X}_i'\mathbf{Y}$.

In-Class Exercise: (Simple linear regression). Consider simple linear regression with the usual model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and also using the no-intercept model $Y_i = \beta_1 x_i + \varepsilon_i$. For both models, find a formula for the estimate of the slope parameter β_1. When will the two estimates be the same?

For the with-intercept model, the least-squares estimate of β_1 is

$$\hat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})Y_i / \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

For the no-intercepet model, the least-squares estimate of β_1 is

$$\hat{\beta}_1 = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2.$$

The slope estimates in the two models will be equal when $\bar{x} = 0$, i.e., when $x = (x_1, \ldots, x_n)'$ is orthogonal to the intercept $\mathbf{1} = (1, \ldots, 1)'$.

In-Class Exercise: Consider an adjustment to the basic linear model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \bar{x} + \beta_1 (x_i - \bar{x}) + \varepsilon_i \\ &= \beta_0^* + \beta_1 (x_i - \bar{x}) + \varepsilon_i, \end{aligned}$$

What is the design matrix? Does it have orthogonal structure? Use the result on page 1 to find $\hat{\beta}_0$ and $\hat{\beta}_1$. Compare your formula for $\hat{\beta}_1$ to the formula for the with-intercept model on the last page.

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix} = (\mathbf{x}_0, \mathbf{x}_1)$$

has orthogonal columns and so by the result on p. 1,

$$\hat{\beta}_0^* = \frac{\mathbf{x}_0' \mathbf{Y}}{\mathbf{x}_0' \mathbf{x}_0} = \bar{Y},$$

$$\hat{\beta}_1 = \frac{\mathbf{x}_1' \mathbf{Y}}{\mathbf{x}_1' \mathbf{x}_1} = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Orthogonality and precision of least-squares estimates:

Theorem: (Seber & Lee Exercise 3e.3). Assume the linear model $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where the columns of $\mathbf{X}$ are linearly independent (so we're in the full rank case). Further suppose that

$$\mathbf{x}_i'\mathbf{x}_i \leq c_i^2,$$

for fixed constants c_i. Then

$$\text{var}(\hat{\beta}_i) \geq \sigma^2/c_i^2,$$

and the minimum is attained when $\mathbf{x}_i'\mathbf{x}_i = c_i^2$, and the columns of $\mathbf{X}$ are orthogonal, i.e., $\mathbf{x}_i'\mathbf{x}_j = 0$, for $j \neq i$.

Example: (2^k factorial design). Suppose that k factors are to be studied to determine their effect on the output of a manufacturing process. Each factor is to be varied within a given plausible range of values and the variables have been scaled so that the range is -1 to $+1$. Then the theorem implies that the optimal design has orthogonal columns and all variables set to $+1$ or -1. Such a design is called a 2^k factorial design.

Lemma: If $\mathbf{A}$ and $\mathbf{D}$ are symmetric and all inverses exist,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{E}^{-1}\mathbf{B}'\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{B}'\mathbf{A}^{-1} & \mathbf{E}^{-1} \end{pmatrix},$$

where $\mathbf{E} = \mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}$.

Proof: Check that the matrix times the candidate inverse matrix gives the identity.

Proof of the Theorem: Let $k = p - 1$. For convenience, assume that β_i is in the last position, i.e., $i = k$, (reordering the columns of $\mathbf{X}$ if necessary).

Write $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k) = (\mathbf{W}, \mathbf{x}_k)$. Then

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{x}_k \\ \mathbf{x}_k'\mathbf{W} & \mathbf{x}_k'\mathbf{x}_k \end{pmatrix}.$$

Now apply the lemma:

$$\operatorname{var}(\hat{\beta}_k) = \frac{\sigma^2}{\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{x}_k}.$$

We can show that $\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{x}_k \le \mathbf{x}_k'\mathbf{x}_k$ It follows that $\operatorname{var}(\hat{\beta}_k) \ge \sigma^2/c_k^2$, with equality if and only if $\mathbf{x}_k'\mathbf{x}_k = c_k^2$ and $\mathbf{x}_j'\mathbf{x}_k = 0$ for all $j \ne k$.

MULTICOLLINEARITY

A useful expression for $\text{var}(\mathbf{a}'\hat{\boldsymbol{\beta}})$ can be obtained from the theorem as follows. Because $\mathbf{X}'\mathbf{X}$ is p.s.d. with rank r, it must have r positive eigenvalues $\lambda_1, \ldots, \lambda_r$ and $p-r$ zero eigenvalues. Therefore, there is an orthogonal $\mathbf{T}$ such that

$$\mathbf{T}'(\mathbf{X}'\mathbf{X})\mathbf{T} = \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0).$$

Now a generalized inverse of $\boldsymbol{\Lambda}$ is (*check it!*)

$$\boldsymbol{\Lambda}^- = \text{diag}(\lambda_1^{-1}, \ldots, \lambda_r^{-1}, 0, \ldots, 0).$$

and a generalized inverse of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^- = \mathbf{T}\boldsymbol{\Lambda}^-\mathbf{T}'.$$

Therefore,

$$\begin{aligned}
\text{var}(\mathbf{a}'\hat{\boldsymbol{\beta}}) &= \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^-\mathbf{a} \\
&= \sigma^2\mathbf{a}'\mathbf{T}\boldsymbol{\Lambda}^-\mathbf{T}'\mathbf{a} \\
&= \sigma^2\mathbf{c}'\boldsymbol{\Lambda}^-\mathbf{c},\ (\mathbf{c} = \mathbf{T}'\mathbf{a}) \\
&= \sigma^2\sum_{i=1}^{r} c_i^2\lambda_i^{-1}.
\end{aligned}$$

This formula identifies the effect of *multicollinearity*, i.e., near linear dependencies in columns of $\mathbf{X}$. Which functions $\mathbf{a}'\boldsymbol{\beta}$ will be most affected by multicollinearity? If $\mathbf{a} = a\boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i$ is a column of $\mathbf{T}$ (eigenvector of $\mathbf{X}'\mathbf{X}$), then $\mathbf{c} = \mathbf{T}'(a\boldsymbol{\alpha}_i) = a\mathbf{e}_i$, where $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ with 1 in the ith component. Then $\text{var}(\mathbf{a}'\hat{\boldsymbol{\beta}}) = a^2\sigma^2\lambda_i^{-1}$. If λ_i is small (but positive), $\text{var}(\mathbf{a}'\hat{\boldsymbol{\beta}})$ will be large.

Suppose we can get one more observation and we can choose the values of the predictor variables $\mathbf{x}'_{n+1} = (x_{n+1,0}, \ldots, x_{n+1,p-1})$ for the next observation Y_{n+1}. What values should we choose?

The model for the full data set is

$$\begin{aligned} \begin{pmatrix} \mathbf{Y} \\ Y_{n+1} \end{pmatrix} &= \begin{pmatrix} \mathbf{X} \\ \mathbf{x}'_{n+1} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\varepsilon} \\ \varepsilon_{n+1} \end{pmatrix} \\ &= \mathbf{X}_* \boldsymbol{\beta} + \boldsymbol{\varepsilon}_*. \end{aligned}$$

Suppose we choose $\mathbf{x}'_{n+1} = a\boldsymbol{\alpha}'_i$ where $\boldsymbol{\alpha}_i$ is an eigenvector of $\mathbf{X}'\mathbf{X}$ with eigenvalue λ_i. Except for this particular eigenvector $\boldsymbol{\alpha}_i$, the eigenvectors of $\mathbf{X}'_*\mathbf{X}_*$ are the same as those of $\mathbf{X}'\mathbf{X}$ with the same eigenvalues:

$$\mathbf{X}_* = \begin{pmatrix} \mathbf{X} \\ \mathbf{x}'_{n+1} \end{pmatrix}$$

$$\mathbf{X}'_*\mathbf{X}_* = \begin{pmatrix} \mathbf{X} & \mathbf{x}_{n+1} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{x}'_{n+1} \end{pmatrix} = \mathbf{X}'\mathbf{X} + \mathbf{x}_{n+1}\mathbf{x}'_{n+1}$$

Let $\boldsymbol{\alpha}$ be an eigenvector of $\mathbf{X}'\mathbf{X}$ (with eigenvalue λ) different from $\mathbf{x}_{n+1}$.

$$\begin{aligned} \mathbf{X}'_*\mathbf{X}_*\boldsymbol{\alpha} &= (\mathbf{X}'\mathbf{X} + \mathbf{x}_{n+1}\mathbf{x}'_{n+1})\boldsymbol{\alpha} = \mathbf{X}'\mathbf{X}\boldsymbol{\alpha} + \mathbf{x}_{n+1}\mathbf{x}'_{n+1}\boldsymbol{\alpha} \\ &= \mathbf{X}'\mathbf{X}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \end{aligned}$$

since the eigenvectors of $\mathbf{X}'\mathbf{X}$ are orthogonal.

For the particular eigenvector $\boldsymbol{\alpha}_i$ with eigenvalue λ_i:

$$\begin{aligned}\mathbf{X}_*'\mathbf{X}_*\boldsymbol{\alpha}_i &= (\mathbf{X}'\mathbf{X} + \mathbf{x}_{n+1}\mathbf{x}_{n+1}')\boldsymbol{\alpha}_i \\ &= \mathbf{X}'\mathbf{X}\boldsymbol{\alpha}_i + a^2\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_i \\ &= \lambda_i\boldsymbol{\alpha}_i + a^2\boldsymbol{\alpha}_i = (\lambda_i + a^2)\boldsymbol{\alpha}_i\end{aligned}$$

Therefore, it may be best to choose $\mathbf{x}_{n+1}$ to be proportional to the eigenvector with the smallest positive eigenvalue. This will increase the smallest positive eigenvalue in the extended design matrix, and thereby decrease the worst-case variability.