**Generalized least squares**. A variance-stabilizing transformation may help in correcting the situation back to the case with (approximately) equal-variance errors, but the transformation applied might, at the same time, impact linearity: if the original means $\mathbb{E}[Y_i]$ are linear in $\boldsymbol{X}_i$, then, by the same delta method argument, after transformation the means are $\mathbb{E}[Y_i] \approx f(\mu_Y)$ and we generally lose linearity.

There is another method to deal with violations of the equal-variance assumption by working directly with the original data, rather than transforming. Thus, assume an *Extended linear model*:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V} \tag{34}$$

where $\boldsymbol{V}$ is a known $n \times n$ positive-definite covariance matrix. Note that in the special case $\boldsymbol{V} = \boldsymbol{I}_n$ we are back to the standard linear model. It is easy to verify that the usual LS estimator,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

is still an unbiased estimator of $\boldsymbol{\beta}$, and, hence, $\hat{\boldsymbol{\theta}} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ is unbiased for $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$. In the special case $\boldsymbol{V} = \boldsymbol{I}_n$, we further have by the Gauss-Markov theorem that $\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ is BLUE, i.e., it has minimum variance among all linear unbiased estimators of $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$. This is no longer true in the case of a general $\boldsymbol{V}$; however, by reducing the model back to the familiar case $\boldsymbol{V} = \boldsymbol{I}_n$, we can obtain a BLUE for the more extended model (11.4): first, we find an invertible $n \times n$ matrix $\boldsymbol{A}$ s.t.

$$\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^\top$$

(this is always possible when $\boldsymbol{V}$ is positive-definite, e.g. we find such $\boldsymbol{A}$ if we orthogonally diagonalize $\boldsymbol{V}$ ). Now define

$$\tilde{\boldsymbol{Y}} = \boldsymbol{A}^{-1}\boldsymbol{Y}, \quad \tilde{\boldsymbol{X}} = \boldsymbol{A}^{-1}\boldsymbol{X}, \quad \tilde{\boldsymbol{\epsilon}} = \boldsymbol{A}^{-1}\boldsymbol{\epsilon}.$$

Then we have

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}, \quad \tilde{\boldsymbol{\epsilon}} \sim \left(\mathbf{0}, \sigma^2 \boldsymbol{I}_n\right), \tag{35}$$

i.e., the usual linear model holds for the transformed variables. The Gauss-Markov theorem then says that the estimator

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\mathrm{GLS}} = \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\right)^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}} &= \left[\boldsymbol{X}^\top \left(\boldsymbol{A}^\top\right)^{-1} \boldsymbol{A}^{-1}\boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \left(\boldsymbol{A}^\top\right)^{-1} \boldsymbol{A}^{-1}\boldsymbol{Y} \\
&= \left[\boldsymbol{X}^\top \left(\boldsymbol{A}\boldsymbol{A}^\top\right)^{-1} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \left(\boldsymbol{A}\boldsymbol{A}^\top\right)^{-1} \boldsymbol{Y} \\
&= \left(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}
\end{aligned}$$

is BLUE for $\boldsymbol{\beta}$ under the original model (34), but this means that this estimator is also BLUE for the transformed model (35) because this is the same $\boldsymbol{\beta}$ in both models. The estimator $\hat{\boldsymbol{\beta}}^{\mathrm{GLS}}$ above is called the generalized least squares (GLS) estimator.

Special case: if $\boldsymbol{V} = \boldsymbol{W} = \operatorname{diag}(w_1, \ldots, w_n)$ is diagonal, meaning that the errors are uncorrelated but do not have equal variances, the GLS estimator is called the weighted least squares (WLS) estimator.

# 9 Multicollinearity

Recall that, throughout, we have assumed that the columns of the $n \times (p+1)$ matrix $\boldsymbol{X}$ are linearly independent (implying necessarily that $p + 1 \leq n$ ). If the columns of $\boldsymbol{X}$ were linearly dependent, then

$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ is not defined because $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible, and there is indeed no unique LS estimator (in that case, any minimizer of the sum of squared residuals is a LS estimate). In fact, even the *true* parameter vector $\boldsymbol{\beta}$ is not well-defined in the sense that it is non-identifiable $\Longleftrightarrow$ there exist several choices of $\boldsymbol{\beta}$ yielding the same value for $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ ).

While we assume that the columns of $\boldsymbol{X}$ are never exactly linearly dependent, i.e.,

$$\boldsymbol{X}\boldsymbol{c} = \sum_{j=0}^{p} c_j \boldsymbol{X}^{(j)} \neq \mathbf{0}$$

for all nonzero $\boldsymbol{c} \in \mathbb{R}^p$, they may still be *nearly* linearly dependent, i.e.,

$$\boldsymbol{X}\boldsymbol{c} = \sum_{j=0}^{p} c_j \boldsymbol{X}^{(j)} \approx \mathbf{0}$$

for some $\boldsymbol{c} \neq \mathbf{0}$. In other words, there is redundancy in the explanatory variables in the sense that there is an explanatory variable that's approximately a linear combination of the others. If this is the case, we will say that there is *multicollinearity* in the $\boldsymbol{X}$ matrix(remark: technically, 'multicollinearity' refers to the case where $\boldsymbol{X}$ has a column that is an exact linear combinations of two or more—hence '*multi*collinearity' rather than just 'collinearity'—the other columns, but here we use the term to describe the case where there's *approximate* multicollinearity). While multicollinearity generally does not affect prediction accuracy (recall that $\hat{\boldsymbol{Y}} = \boldsymbol{P}_{\mathrm{Im}(\boldsymbol{X})}\boldsymbol{Y}$ does not depend on $\boldsymbol{X}$ itself, only on the span of its columns), it does affect the variance of the coefficients of the explanatory variables. Specifically, if there's substantial multicollinearity, the LS estimator $\hat{\boldsymbol{\beta}}$ will be highly sensitive to small changes in $\boldsymbol{Y}$, which will result in large variances for the estimators $\hat{\beta}_j$. We first explain why this is the case, by giving an alternative representation for the LS coefficient $\hat{\beta}_j$.

**Obtaining the LS estimates $\hat{\beta}_j$ through simple regression**. The formula $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{Y}$ gives the entire $(p+1)$-dimensional vector of LS estimates $\hat{\beta}_j,\ j = 0, 1, ..., p$ at once; if we want to obtain the estimate for an individual coefficient $\beta_j$, we can simply extract the $j$-th element of $\hat{\boldsymbol{\beta}}$,

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

which is equivalent to $\hat{\beta}_j = \boldsymbol{e}_j^\top \hat{\boldsymbol{\beta}}$, and requires calculating the full vector estimate $\hat{\boldsymbol{\beta}}$ first. We now present an alternative way to calculate $\hat{\beta}_j$ through a *simple* regression. Remember that calculating the LS solution for the simple regression of $\mathbf{Y}$ on $\boldsymbol{X}^{(j)}$, the $j$-th column of $\boldsymbol{X}$, will give an estimate of the coefficient of the $j$-th predictor in the model that includes *only* the $j$-th predictor (and an intercept),

$$Y_i = \beta_0^* + \beta_j^* X_{ij} + \epsilon_i^*, \tag{36}$$

whereas the LS estimate $\hat{\beta}_j$ from the *multiple* regression of $\mathbf{Y}$ on $\boldsymbol{X}$ estimates the coefficient of the $j$-th predictor in the model that includes the $j$-th predictor *along with* the other $p-1$ predictors (columns of $\boldsymbol{X}$),

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_j X_{ij} + \beta_p X_{ip} + \epsilon_i. \tag{37}$$

We used different notation, $\beta_j^*$ and $\beta_j$, because these parameters are indeed different in general, and they have different interpretations: $\beta_j^*$ is the increase in the mean value of $Y$ when $X_j$ increases by one unit, whereas $\beta_j$ is the increase in the mean value of $Y$ when $X_j$ increases by one unit and the other predictors *are held* *fixed* (basically, *conditional* on the values of the remaining $p-1$ predictors). Note that the errors are also not the same, which is why we used different notation $\epsilon_i^*$ vs $\epsilon_i$.

While fitting a simple regression of $\mathbf{Y}$ on $\boldsymbol{X}^{(j)}$ gives $\hat{\beta}_j^*$, which is *not* what we want, the required estimate $\hat{\beta}_j$ can in fact be obtained by simple regression (without intercept) on an *adjusted* version of $\boldsymbol{X}^{(j)}$. Specifically, do the following:

1. Step 1: regress $\boldsymbol{X}^{(j)}$ on $\boldsymbol{X}^{(-j)}$, the $n \times p$ matrix obtained by deleting the $j$-th column from $\boldsymbol{X}$.

2. Step 2: calculate the residuals for the regression in Step 1,

$$\tilde{\boldsymbol{X}}^{(j)} = (\boldsymbol{I}_n - \boldsymbol{P}_{-j})\boldsymbol{X}^{(j)},$$

   where $\boldsymbol{P}_{-j}$ is the projection matrix onto the image of $\boldsymbol{X}^{(-j)}$ (remark: $\tilde{\boldsymbol{X}}^{(j)}$ is sometimes referred to as the *adjustment* of $\boldsymbol{X}^{(j)}$ to the other predictors in the model, being the projection of $\boldsymbol{X}^{(j)}$ to the orthogonal complement of $\mathrm{Im}(\boldsymbol{X}^{(-j)})$).

3. Step 3: fit a simple regression *without intercept* of $\mathbf{Y}$ on $\tilde{\boldsymbol{X}}^{(j)}$,

$$\hat{\gamma}_j := \operatorname*{arg\,min}_{c\in\mathbb{R}} \|\mathbf{Y} - c\tilde{\boldsymbol{X}}^{(j)}\|^2 = \frac{\tilde{\boldsymbol{X}}^{(j)\top}\boldsymbol{Y}}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2} \in \mathbb{R} \tag{38}$$

   (remark: recall that for simple regression *with* intercept, the LS estimator would minimize $\|\mathbf{Y} - b_0 - b_1\tilde{\boldsymbol{X}}^{(j)}\|^2$ over $b_0, b_1$).

**Proposition 9.** *For the algorithm described above, we have $\hat{\beta}_j = \hat{\gamma}_j$.*
*(in words: the LS coefficient $\hat{\beta}_j$ in the multiple regression of $\mathbf{Y}$ on $\boldsymbol{X}$, is exactly equal to the LS coefficient in the* simple *regression, without intercept, of $\mathbf{Y}$ on $\tilde{\boldsymbol{X}}^{(j)}$, the residual from regressing the $j$th predictor $\boldsymbol{X}^{(j)}$ on the remaining predictors $\boldsymbol{X}^{(-j)}$).*

*Proof.* First recall the general fact (used in (38) above) that the projection of $\boldsymbol{v} \in \mathbb{R}$ on $\boldsymbol{u} \in \mathbb{R}$ is given by $\alpha\boldsymbol{u}$ where $\alpha = \frac{\langle \boldsymbol{v},\boldsymbol{u}\rangle}{\|u\|^2}$. Now, the projection of $\mathbf{Y}$ on $\tilde{\boldsymbol{X}}^{(j)}$ is the same as the projection of $\hat{\mathbf{Y}}$ on $\tilde{\boldsymbol{X}}^{(j)}$, because $\mathbf{Y} = \hat{\mathbf{Y}} + \boldsymbol{e}$, and $\boldsymbol{e} \perp \tilde{\boldsymbol{X}}^{(j)}$ (since $\tilde{\boldsymbol{X}}^{(j)}$ is still in the image of $\boldsymbol{X}$). Write $\hat{\mathbf{Y}} = \sum_{k=0}^p \hat{\beta}_k \boldsymbol{X}^{(k)}$ where $\hat{\beta}_0, ..., \hat{\beta}_p$ are the LS estimates in the multiple regression of $\mathbf{Y}$ on $\boldsymbol{X}$. Then if $\hat{\gamma}_j\tilde{\boldsymbol{X}}^{(j)}$ is the projection of $\hat{\mathbf{Y}}$ on $\tilde{\boldsymbol{X}}^{(j)}$, by the general fact mentioned in the beginning we have

$$\hat{\gamma}_j = \frac{\langle \sum_{k=0}^p \hat{\beta}_k \boldsymbol{X}^{(k)}, \tilde{\boldsymbol{X}}^{(j)}\rangle}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2} = \frac{\sum_{k=0}^p \hat{\beta}_k \langle \boldsymbol{X}^{(k)}, \tilde{\boldsymbol{X}}^{(j)}\rangle}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2} \overset{(i)}{=} \frac{\hat{\beta}_j \langle \boldsymbol{X}^{(j)}, \tilde{\boldsymbol{X}}^{(j)}\rangle}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2} \overset{(ii)}{=} \frac{\hat{\beta}_j \langle \tilde{\boldsymbol{X}}^{(j)}, \tilde{\boldsymbol{X}}^{(j)}\rangle}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2} = \hat{\beta}_j$$

where in $(i)$ we used the fact that $\tilde{\boldsymbol{X}}^{(j)}$ is orthogonal to all $\boldsymbol{X}^{(k)},\ k \neq j$, and $(ii)$ is because $\langle \boldsymbol{X}^{(j)}, \tilde{\boldsymbol{X}}^{(j)}\rangle = \langle \tilde{\boldsymbol{X}}^{(j)} + \boldsymbol{z}, \tilde{\boldsymbol{X}}^{(j)}\rangle$ where $\boldsymbol{z}$ is a linear combination of $\boldsymbol{X}^{(k)},\ k \neq j$. $\square$

Using Proposition 9, we can calculate the variance of $\hat{\beta}_j$ (under the original linear model, $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$) as

$$\operatorname{Var}\left(\hat{\beta}_j\right) = \operatorname{Var}\left(\hat{\gamma}_j\right) = \operatorname{Var}\left(\frac{\boldsymbol{e}^{(j)\top}\boldsymbol{Y}}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2}\right) = \operatorname{cov}\left(\frac{\boldsymbol{e}^{(j)\top}\boldsymbol{Y}}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2}\right) = \sigma^2 \frac{\boldsymbol{e}^{(j)\top}\tilde{\boldsymbol{X}}^{(j)}}{\left(\boldsymbol{e}^{(j)\top}\tilde{\boldsymbol{X}}^{(j)}\right)^2} = \sigma^2 \frac{1}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2}$$

(Remark: note that this means $\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{jj} = \frac{1}{\|\tilde{\boldsymbol{X}}^{(j)}\|^2}\Bigg)$.

Now, returning to the discussion on multicollinearity, if the columns of $\boldsymbol{X}$ are nearly linearly dependent, this means that the squared norm of $\tilde{\boldsymbol{X}}^{(j)} = (\boldsymbol{I}_n - \boldsymbol{P}_{-j})\boldsymbol{X}^{(j)}$ will have small norm. This suggests that $\operatorname{Var}(\hat{\beta}_j)$ will be large.

**Basic checks for multicollinearity**.

1. Look at the Pearson correlations (pairwise correlations) for all pairs of explanatory variables. High absolute values are a sign of redundancy.

2. For each $j = 0, 1, \ldots, p$, look at the $R^2$ value in the regression of the $j$-the explanatory variable $\boldsymbol{X}^{(j)} = (X_{1j}, \ldots, X_{nj})^\top$ on the remaining $p$ explanatory variables $\boldsymbol{X}^{(-j)}$. If we denote this by

$$R_j^2 := \frac{SSR_j}{SST_j}$$

where $SST_j$ and $SSR_j$ are the SST and SSR in the multiple regression of $\boldsymbol{X}^{(j)}$ on $\boldsymbol{X}^{(-j)}$, then large values of $R_j^2$ means that $\boldsymbol{X}^{(j)}$ can be approximated with high accuracy as a linear combination of the others, ie., the residuals of the simple regression of $\mathbf{Y}$ on $\tilde{\boldsymbol{X}}^{(j)}$ are small. Two standard metrics related functionally to $R_j^2$ are the *Tolerance* and the *Variance Inflation Factor* (VIF),

$$\operatorname{Tol}_j := 1 - R_j^2 \quad \text{and} \quad \operatorname{VIF}_j := \frac{1}{1 - R_j^2}.$$

Hence, small values of $\operatorname{Tol}_j$, or large values of $\operatorname{VIF}_j$, are indication for a problem (redundancy). As a rough guideline, $R_j^2$ values exceeding 0.85 , i.e. $\operatorname{Tol}_j$ falling below 0.15 or $\operatorname{VIF}_j$ exceeding 6.6, can be considered extreme (indicating substantial multicollinearity).

*Remark*: The Variance Inflation Factor gets its name from the following fact: let $\hat{\beta}_j^*$ denote the LS estimate in a simple regression (with intercept) of $\mathbf{Y}$ on $\boldsymbol{X}^{(j)}$, and, as usual, $\hat{beta}_j$ is the LS estimate for the $j$th predictor in the multiple regression of $\mathbf{Y}$ on $\boldsymbol{X}$. Then we know that

$$\operatorname{Var}(\hat{\beta}_j^*) = \sigma^2 F_*^{-1}, \quad F_* = \sum_{i=1}^n (X_{ij} - \bar{X}_{\cdot j})^2 = SST_j,$$

and

$$\operatorname{Var}(\hat{\beta}_j) = \sigma^2 F^{-1}, \quad F = \|\boldsymbol{e}^{(j)}\|^2 = SSE_j,$$

Therefore, the ratio of these variances is

$$\frac{\operatorname{Var}(\hat{\beta}_j)}{\operatorname{Var}(\hat{\beta}_j^*)} = \frac{F^*}{F} = \frac{SST_j}{SSE_j} = (1 - R_j^2)^{-1},$$

which is exactly the definition of $VIF_j$. Note also that $VIF_j \geq 1$ (because $R_j^2 \leq 1$), so we conclude that the variance of the LS estimate of the $j$th predictor necessarily inflates when moving from the simple regression of $\mathbf{Y}$ on $\boldsymbol{X}^{(j)}$ to the multiple regression of $\mathbf{Y}$ on $\boldsymbol{X}$.

3. Condition number and condition index. The *condition number* of a matrix $\boldsymbol{X}$ (whose columns are linearly independent) is defined by

$$\gamma(\boldsymbol{X}) := \frac{\max_{\|\boldsymbol{c}\|=1} \|\boldsymbol{X}\boldsymbol{c}\|}{\min_{\|\boldsymbol{c}\|=1} \|\boldsymbol{X}\boldsymbol{c}\|} \tag{39}$$

Large values of $\gamma(\boldsymbol{X})$ indicate higher degree of redundancy (the restriction $\|\boldsymbol{c}\| = 1$ keeps the numerator and denominator calibrated); indeed, in that case the denominator is approximately zero. The smallest possible value for $\gamma(\boldsymbol{X})$ is 1, which obtains when the column of $\boldsymbol{X}$ are orthogonal with the same norm, i.e., $\boldsymbol{X}^\top \boldsymbol{X} \propto \boldsymbol{I}_{p+1}$. As a rough guideline, we can consider values of $\gamma(\boldsymbol{X})$ between 5-10 as low degree of multicollinearity, and between $30 - 100$ as high degree of multicollinearity. It can be shown, by considering the diagonal representation of the positive-definite matrix $\boldsymbol{X}^\top \boldsymbol{X}$, that the numerator in (39) is equal to the square root of the largest eigenvalue of $\boldsymbol{X}^\top \boldsymbol{X}$, and the denominator in (39) to the the square root of the smallest eigenvalue, so that

$$\gamma(\boldsymbol{X}) := \left( \frac{\lambda_{\max}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)}{\lambda_{\min}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)} \right)^{1/2}$$

More generally, we we define the condition index corresponding to the $j$-th eigenvalue $\lambda_j$ to be

$$\alpha_j := \left( \frac{\lambda_{\max}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)}{\lambda_j\left(\boldsymbol{X}^\top \boldsymbol{X}\right)} \right)^{1/2}.$$

Since $\lambda_j = \|\boldsymbol{X}\boldsymbol{u}_j\|$ where $\boldsymbol{u}_j$ is a unit vector in the direction of the $j$-th eigenvalue (equivalently, the $j$ th column of a matrix $\boldsymbol{U}$ holding an orthonormal diagonalizing basis), small values of $\alpha_j$ indicate "directions" with substantial redundancy, and we can identify explanatory variables $\boldsymbol{X}^{(j)}$ exhibiting redundancy by looking at the entries of the corresponding $\boldsymbol{u}_j$ that have the largest coefficients.

For example, in a case with 3 explanatory variables (+intercept) the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ are

$$400.0565138 \quad 200.0000000 \quad 199.7850065 \quad 0.1584797$$

with corresponding condition indices

$$1.000000 \quad 1.414313 \quad 1.415074 \quad 50.242803,$$

then $\lambda_{\max}$ is very large compared to $\lambda_{\min}$, indicating near linear dependence in the linear combination corresponding to the eigenvector of the smallest eigenvalue. The eigenvectors are

```
> eig$vectors
            [,1] [,2]        [,3]         [,4]
[1,]  0.00000000    1  0.00000000  0.000000000
[2,] -0.70670271    0  0.02461521  0.707082292
[3,] -0.70674878    0  0.02180557 -0.707128479
[4,]  0.03282445    0  0.99945916 -0.001986711
```

The last eigenvector (corresponding to $\lambda_{\min}$ ) is the problematic "direction", and we see that it is approximately equal to $.707\boldsymbol{X}^{(1)} - .707\boldsymbol{X}^{(2)}$, equivalently to $\boldsymbol{X}^{(1)} - \boldsymbol{X}^{(2)}$, indicating that $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are nearly linearly dependent.

4. Proportion of variance table. Recall that

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

Now, if $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ is a spectral decomposition of $\boldsymbol{X}^\top \boldsymbol{X}$, then $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top$ is a spectral decomposition of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and we have

$$\operatorname{Var}\left(\hat{\beta}_j\right) = [\operatorname{cov}(\hat{\boldsymbol{\beta}})]_{jj} = \sum_r \lambda_r^{-1} \boldsymbol{u}_j \boldsymbol{u}_j^\top = \sum_r \lambda_r^{-1} U_{jr}^2$$

The quantity

$$\Pi_{rj} = \frac{\lambda_r^{-1} U_{jr}^2}{\sum_s \lambda_s^{-1} U_{js}^2}$$

is the *proportion of* $\operatorname{Var}\left(\hat{\beta}_j\right)$ contributed by the "direction" (=linear combination of the original explanatory variables $\left.\boldsymbol{X}^{(j)}\right)$ corresponding to $\lambda_r$, i.e., that represented by the eigenvector $\boldsymbol{u}_r$. For a small value $\lambda_r$, we can identify "problematic" combinations by finding $j$ 's with large value of $\Pi_{rj}$.

In R, Tol, VIF and variance proportions can be calculated automatically using the function `ols_coll_diag` in the R package `olsr`. If I understand correctly, the package first normalizes all explanatory variables so that the diagonal entries of $\boldsymbol{X}^\top \boldsymbol{X}$ are all 1's, then diagonalizes the resulting matrix.

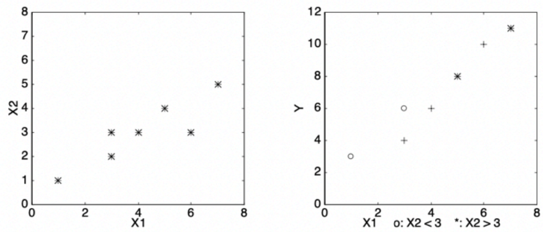
*Example*.

```
> coll.ans = ols_coll_diag(mdl1)
> coll.ans
Tolerance and Variance Inflation Factor
---------------------------------------
# A tibble: 3 x 3
Variables Tolerance    VIF
<chr>         <dbl>  <dbl>
1 x1s         0.00158 631.
2 x2s         0.00158 631.
3 x3s         0.995     1.01
```

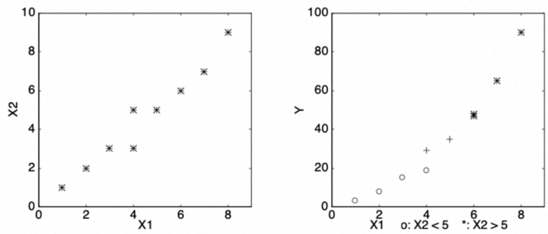
```
Eigenvalue and Condition Index
------------------------------
Eigenvalue Condition Index intercept          x1s          x2s          x3s
1 2.0002825690        1.000000         0 3.955611e-04 3.955611e-04 0.0005356927
2 1.0000000000        1.414313         1 0.000000e+00 0.000000e+00 0.0000000000
3 0.9989250326        1.415074         0 9.609606e-07 7.540096e-07 0.9945105139
4 0.0007923985       50.242803         0 9.996035e-01 9.996037e-01 0.0049537934
```

We see that almost $100\%$ of the variance in $\hat{\beta}_1$ and $\hat{\beta}_2$ come from the term with the low eigenvalue, thus indicating a multicollinearity problem.

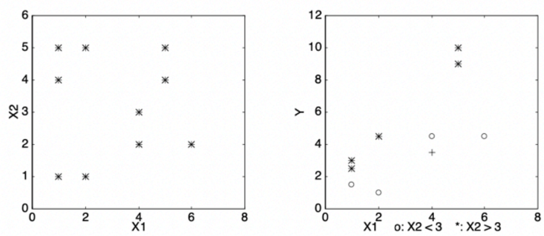
**Visual checks for multicollinearity and interactions**. We give some illustrating examples for how multicollinearity and interaction each look visually in graphs (credit to Prof. Sam Oman).



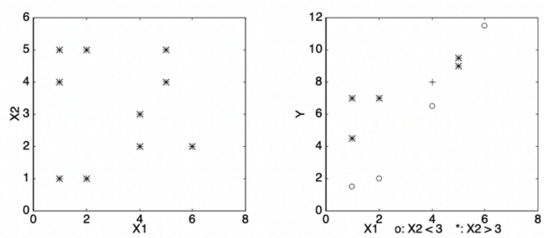

**Set A**: multicollinearity, no interaction




**Set B**: multicollinearity, interaction




**Set C**: no multicollinearity, interaction




**Set D**: no multicollinearity, no interaction

Note: Sets C, D have the same values.