Linear transformations of regressors.

Transformations of regressors

One–hot encodings and constants

Recall in the Ames housing data, we ran the following two regressions:

$$y_n \sim \beta_e x_{ne} + \beta_g x_{ng}$$
$$y_n \sim \gamma_0 + \gamma_g x_{ng} + \varepsilon_n = z_n^\top \gamma,$$

where I take $\gamma = (\gamma_0, \gamma_g)^\top$ and $z_n = (1, x_{ng})^\top$.

We found using `R` that the best fits were given by

$$\begin{aligned} \hat{\beta}_e &= \bar{y}_e & \hat{\beta}_g &= \bar{y}_g \\ \hat{\gamma}_0 &= \bar{y}_e & \hat{\gamma}_g &= \bar{y}_g - \bar{y}_e \end{aligned}$$

We can compute the latter by constructing the Z matrix whose rows are $z_n^\top$. (We use Z to differentiate the X matrix from the previous example.) Using similar reasoning to the one–hot encoding, we see that

$$Z^\top Z = \begin{pmatrix} N & N_g \\ N_g & N_g \end{pmatrix}.$$

This is invertible as long as $N_g \neq N$, i.e., as long as there is at least one $k_n = e$. We have

$$(Z^\top Z)^{-1} = \frac{1}{N_g(N - N_g)} \begin{pmatrix} N_g & -N_g \\ -N_g & N \end{pmatrix} \quad \text{and} \quad Z^\top Y = \begin{pmatrix} \sum_{n=1}^{N} y_n \\ \sum_{n:k_n=g} y_n \end{pmatrix}$$

It is possible (but a little tedious) to prove $\hat{\gamma}_0 = \bar{y}_e$ and $\hat{\gamma}_g = \bar{y}_g - \bar{y}_e$ using these formulas. But an easier way to see it is as follows.

Note that $x_{ne} + x_{ng} = 1$. That means we can always re-write the regression with a constant as

$$y_n \sim \gamma_0 + \gamma_g x_{ng} = \gamma_0(x_{ne} + x_{ng}) + \gamma_g x_{ng} = \gamma_0 x_{ne} + (\gamma_0 + \gamma_g) x_{ng}.$$

Now, we already know from the one–hot encoding case that the sum of squared residuals is minimized by setting $\hat{\gamma}_0 = \bar{y}_e$ and $\hat{\gamma}_0 + \hat{\gamma}_g = \bar{y}_g$. We can then solve for $\hat{\gamma}_g = \bar{y}_g - \bar{y}_e$, as expected.

This is case where we have two regressions whose regressors are invertible linear combinations of one another:

$$z_n = \begin{pmatrix} 1 \\ x_{ng} \end{pmatrix} = \begin{pmatrix} x_{ne} + x_{ng} \\ x_{ng} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_{ng} \\ x_{ne} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} x_n.$$

It follows that if you can acheive a least squares fit with $x_n^\top \hat{\beta}$, you can achieve exactly the same fit with

$$\hat{\beta}^\top x_n = \hat{\beta}^\top \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^{-1} z_n,$$

which can be achieved by taking

$$\hat{\gamma}^\top = \hat{\beta}^\top \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^{-1} \Rightarrow \hat{\gamma} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^{-T} \hat{\beta} = \frac{1}{-1} \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix} \hat{\beta} = \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_1 - \hat{\beta}_2 \end{pmatrix},$$

exactly as expected.

We will see this is an entirely general result: when regressions are related by invertible linear transformations of regressors, the fit does not change, but the optimal coefficients are linear transforms of one another.

Redundant regressors

Suppose we run the (silly) regression $y \sim \alpha \cdot 1 + \gamma \cdot 3 + \varepsilon_n$. That is, we regress on both the constant 1 and the constant 3. We have

$$X = \begin{pmatrix} 1 & 3 \\ 1 & 3 \\ 1 & 3 \\ \vdots & \end{pmatrix} = (1 \quad 31)$$

and so

$$X^\top X = \begin{pmatrix} 1^\top 1 & 31^\top 1 \\ 31^\top 1 & 91^\top 1 \end{pmatrix} = N \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix}$$

This is not invertible (the second row is 3 times the first, and the determinant is $9 - 3 \cdot 3 = 0$). So $\hat{\beta}$ is not defined. What went wrong?

One way to see this is to define $\beta = \alpha + 3\gamma$ and write

$$y_n = (\alpha + 3\gamma) + \varepsilon_n = \beta + \varepsilon_n.$$

There is obviously only one $\hat{\beta}$ that minimizes $\sum_{n=1}^N \varepsilon_n^2$, $\hat{\beta} = \bar{y}$. But there are an infinite set of choices for α and γ satisfying

$$\alpha + 3\gamma = \hat{\beta} = \bar{y}.$$

Specifically, for any value of γ we can take $\alpha = \bar{y} - 3\gamma$, leaving β unchanged. All of these choices for α, γ acheive the same $\sum_{n=1}^N \varepsilon_n^2$! So the least squares criterion cannot distinguish among them.

In general, this is what it means for $X^\top X$ to be non–invertibile. It happens precisely when there are redundant regressors, and many regression coefficients that result in the same fit.

Redundant regressors and zero eigenvalues

In fact, $X^\top X$ is invertible precisely when $X^\top X$ has a zero eigenvalue. In the preceding example, we can see that

$$X^\top X \begin{pmatrix} 3 \\ -1 \end{pmatrix} = N \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

so $(3, -1)^\top$ is a zero eigenvector. (In general you might find this by numerical eigenvalue decomposition, but in this case you can just guess the zero eigenvalue.)

Going back to **?@eq-ols-esteq**, we see that this means that

$$X^\top Y = (X^\top X) \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix} = N \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix} = N \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix} \left(\begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix} + C \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right)$$

for *any* value of C. This means there are an infinite set of "optimal" values, all of which set the gradient of the loss to zero, and all of which have the same value of the loss function (i.e. acheive the same fit). And you can check that these family of values are exactly the ones that satisfy $\alpha + 3\gamma = \hat{\beta} = \bar{y}$, since

$$\alpha + 3\gamma = \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} = 0.$$

Soon, we will see that this is a general result: when $X^\top X$ is not invertible, that means there are many equivalent least squares fits, all characterized precisely by the zero eigenvectors of $X^\top X$.

Zero variance regressors

An example of redundant regressors occurs when the sample variance of x_n is zero and a constant is included in the regression. Specifically, suppose that $\overline{xx} - \overline{x}^2 = 0$.

> **Exercise**
>
> Prove that $\overline{xx} - \overline{x}^2 = 0$ means x_n is a constant with $x_n = \bar{x}$. Hint: look at the sample variance of x_n.

Let's regress $y_n \sim \beta_1 + \beta_2 x_n$.

For simplicity, let's take $x_n = 3$. In that case we can rewrite our estimating equation as

$$y_n = \beta_1 + \beta_2 x_n + \varepsilon_n = (\beta_1 + \beta_2 \bar{x}) + \varepsilon_n.$$

We're thus in the previous setting with $\bar{x}$ in place of the number 3.

Orthogonal regressors

Suppose we have regressors such that the columns of X are orthonormal. This seems strange at first, since we usually specify the *rows* of the regressors, not the columns. But in fact we have seen a near–example with one–hot encodings, which are defined row–wise, but which produce orthogonal column vectors in X. If we divide a one–hot encoding by the square root of the number of ones in the whole dataset, we produce an normal column vector.

If X has orthonormal columns, then $X^\top X = I$, the identity matrix, and so

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = X^\top Y.$$

This is of course the same answer we would have gotten if we had tried to write Y in the basis of the column vectors of X:

$$\begin{aligned} Y &= \hat{\beta}_1 X_{\cdot 1} + ... + \hat{\beta}_P X_{\cdot P} = X\hat{\beta} \Rightarrow \\ X^\top Y &= X^\top X\hat{\beta} = \hat{\beta} \end{aligned}$$

This regression is particularly simple — each component of $\hat{\beta}$ depends only on its corresponding column of X.

Note that if each entry of x_n is mean zero, unit variance, and uncorrelated with the other entries, then $\frac{1}{N}X^\top X \to I$ by the LLN. Such a regressor matrix is not typically orthogonal for any particular N, but it approaches orthogonality as N grows.