Linear Regression

Gianluca Cubadda
Università di Roma "Tor Vergata"

2nd February 2018

The Multiple Regression Model

- The basic assumptions of the multiple regression model are that:

$$\mathrm{E}(Y|X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j,$$

  and the conditional distribution $Y|X$ has finite 2^{nd} and 4^{th} moments.
- Suppose to have N observations of both the target variable Y and each of the predictor X_j in the training set. We can write the model in the matrix format:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

  where $\mathbf{y}$ is the $N-$vector of outputs, $\mathbf{X}$ is the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position), ϵ is the $N-$vector of (unobservable) errors, and $\beta = [\beta_0, ..., \beta_p]'$.

Ordinary Least Squares

- We can get a consistent estimate of the coefficient vector β by minimizing the following quadratic loss function:

  $$\mathrm{RSS}(b) = (\mathbf{y} - \mathbf{X}b)'(\mathbf{y} - \mathbf{X}b)$$

  where b is a generic $(p+1)-$vector.

- The solution of this minimization problem is the Ordinary Least Squares (OLS) estimator:

  $$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- The fitted values of the training inputs are

  $$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \mathbf{H}\mathbf{y}$$

  where the "hat" matrix $\mathbf{H}$ is equal to $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

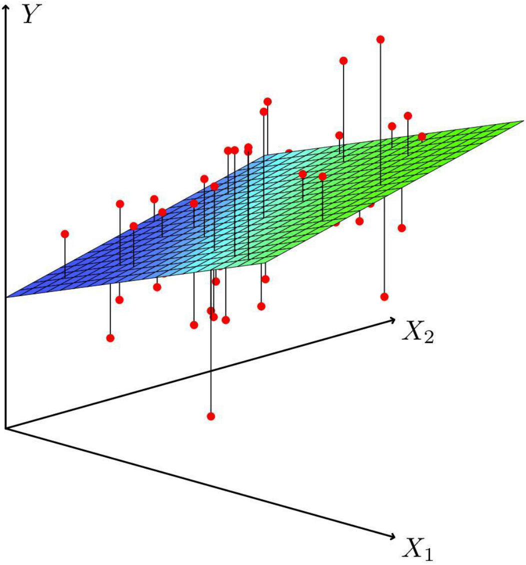
OLS (cont'd)

- The residuals are defined as

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I}_N - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y}$$

  where $\mathbf{M} =(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$.
- Geometric interpretation: the matrices $\mathbf{H}$ and $\mathbf{M}$ respectively project the input vector $\mathbf{y}$ into the space spanned by the columns of $\mathbf{X}$ and into its null space.
- Algebraic implication: the fitted value and the residual vectors are orthogonal each other, that is $\widehat{\mathbf{y}}'\mathbf{e} = 0$.
- Note: $\widehat{\beta}$ cannot be computed when the matrix $\mathbf{X}'\mathbf{X}$ is not invertible. This occurs when $p > N$. Moreover, numerical problems in inverting $\mathbf{X}'\mathbf{X}$ arise when either p approaches N or the input variables are highly correlated each other.

OLS (cont'd)



Statistical properties

- Under the previously given assumptions and treating $\mathbf{X}$ as fixed, $\widehat{\beta}$ is an unbiased and consistent estimator of β , that is

$$\mathrm{E}(\widehat{\beta}) = \beta$$

$$\widehat{\beta} \xrightarrow{p} \beta \text{ as } N \rightarrow \infty$$

- If we additionally assume that $\mathrm{Var}(\epsilon) = \sigma^2 \mathbf{I}_N$ then

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} N\left(0, \sigma^2 S_{XX}^{-1}\right) \text{ where } \mathbf{X}'\mathbf{X}/N \rightarrow S_{XX}$$

- If we further assume that $\epsilon \sim N(0, \sigma^2 \mathbf{I}_N)$ then

$$\widehat{\beta} \sim N\left(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

Moreover, $\widehat{\beta}$ is the maximum likelihood estimator of β.

Measures of fit

- We compute the Residual Standard Error

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{N-p-1}} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{N-p-1}}$$

  where RSS denotes the Residual Sum of Squares.

- R-squared or fraction of variance explained is

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{(\mathbf{y} - \overline{y}1_N)'(\mathbf{y} - \overline{y}1_N)}$$

  where $\overline{y}$ is the average of Y in the training sample, 1_N is a $N-$vector of ones and TSS denotes the Total Sum of Squares.

- Since R^2 is a non decreasing function of p, when comparing models with a different number of predictors we use the adjusted R-squared

$$R_A^2 = 1 - \frac{\mathrm{RSS}/(N-p-1)}{\mathrm{TSS}/(N-1)}$$

Testing

- To test the hypothesis that $\beta_j = 0$, we use the $t-$test statistics or $Z-$scores

$$z_i = \frac{\widehat{\beta}_j}{\mathrm{RSE}\sqrt{v_j}}$$

where v_j is $j-$th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Under H_0, we have $z_i \xrightarrow{d} N(0,1)$ and, under normality, $z_i \sim t_{N-p-1}$.

- To test for the significance of $q(\leq p)$ coefficients simultaneously, we use the $F-$statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(N-p-1)}$$

where $\mathrm{RSS}_0 = \mathbf{e}_0'\mathbf{e}_0$, and $\mathbf{e}_0$ is the residual vector of the model with $(p-q)$ predictors. Under H_0, $qF \xrightarrow{d} \varkappa^2(q)$ and, under normality, $F \sim F(q, N-p-1)$.

- Heteroskedasticity and autocorrelation robust versions of these tests should be used when $\mathrm{Var}(\epsilon) \neq \sigma^2\mathbf{I}_N$.

Ridge

- The ridge coefficients minimize a penalized residual sum of squares

$$(\mathbf{y} - \mathbf{X}b)'(\mathbf{y} - \mathbf{X}b) + \lambda b'b = \mathrm{RSS}(b) + \lambda b'b$$

  where $\lambda (> 0)$ is defined as the tuning parameter.
- The solution of this minimization problem is the ridge estimator:

$$\widehat{\beta}_r = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$$

- Note: given that the ridge solution is not invariant to changes of scale and that there is no need to shrink the intercept term, the inputs are demeaned and standardized and, consequently, β_0 is dropped from β in applications.
- Note: $\widehat{\beta}_r$ can be computed even when $\mathbf{X}'\mathbf{X}$ is singular!

Ridge (cont'd)

- Under the classical assumptions on linear regression, $\widehat{\beta}_r$ is a biased but consistent estimator of β , that is

$$\mathrm{E}(\widehat{\beta}_r) = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}\beta$$

$$\widehat{\beta}_r \xrightarrow{p} \beta \text{ as } N \rightarrow \infty$$

- Moreover, we have

$$\mathrm{Var}(\widehat{\beta}_r) = \sigma^2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

- Notice: there exists a λ such that $\mathrm{MSE}(\widehat{\beta}_r) < \mathrm{MSE}(\widehat{\beta})$. However, the optimal choice of λ depends on unknown parameters.
- Assuming Gaussian errors, $t-$tests and $F-$tests can be used on the ridge estimator (formulae are messy!). They have the same distributions as in the OLS case.

Orthogonal inputs

- The eigen-decomposition of $\mathbf{X}'\mathbf{X}$ is

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

where $\mathbf{D}^2$ is the diagonal $p \times p-$matrix of the (ordered) eigenvalues $(d_1^2 \geq d_2^2 \geq ... \geq d_p^2)$, and $\mathbf{V}$ is $p \times p-$matrix with columns being the eigenvectors such that $\mathbf{V}\mathbf{V}' = \mathbf{I}_p$.

- The principal components of $\mathbf{X}$ are defined as

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

and their variance-covariance matrix (remember that the $\mathbf{X}'s$ are centered, so the $\mathbf{Z}'s$ are) is

$$\mathbf{Z}'\mathbf{Z}/N = \mathbf{D}^2/N$$

- Interpretation: The $j-$th principal component $\mathbf{z}_j = \mathbf{X}v_j$, where v_j is the $j-$th eigenvector, has variance d_j^2/N, subject to being orthogonal to the other ones. The $1-$st principal component $\mathbf{z}_1$ has the largest variance amongst all normalized linear combinations of the $\mathbf{X}'s$.

Orthogonal inputs (cont'd)

- We can rewrite the model into is canonical form

$$\mathbf{y} = \mathbf{Z}\theta + \epsilon$$

  where $\theta = \mathbf{V}'\beta$. Notice that the inputs in $\mathbf{Z}$ are uncorrelated each other.

- The OLS of estimator of θ is

$$\widehat{\theta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{D}^{-2}\mathbf{Z}'\mathbf{y}$$

  whereas the ridge estimator is

$$\widehat{\theta}_r = (\mathbf{Z}'\mathbf{Z}+\lambda\mathbf{I}_p)^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{D}^2+\lambda\mathbf{I}_p)^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{D}^2+\lambda\mathbf{I}_p)^{-1}\mathbf{D}^2\widehat{\theta}$$

Orthogonal inputs (cont'd)

- Hence, the $j-$th element of $\widehat{\theta}_r$ is given by
$$\widehat{\theta}_{r,j} = \widehat{\theta}_j d_j^2/(d_j^2 + \lambda).$$
where $\widehat{\theta}_j$ is the $j-$th element of $\widehat{\theta}$.
- Interpretation: $\widehat{\theta}_{r,j}$ shrinks $\widehat{\theta}_j$ toward 0. Moreover, ridge regression shrinks low-variance principal components (small d_j) more than it does high-variance ones (large d_j).
- Note: the conclusions above hold for the canonical form. We can transform back by setting
$$\widehat{\beta}_r = \mathbf{V}\widehat{\theta}_r$$
but $\widehat{\beta}_r$ may not shrink every component of $\widehat{\beta}$. However, what we can say is that
$$\widehat{\beta}_r'\widehat{\beta}_r = \widehat{\theta}_r'\widehat{\theta}_r < \widehat{\theta}'\widehat{\theta} = \widehat{\beta}'\widehat{\beta}$$
for $\lambda > 0$, so that $\widehat{\beta}_r$ is a shrinkage estimator.

Lasso

- The Lasso (Least Absolute Shrinkage and Selection Operator) minimizes a penalized residual sum of squares

$$(\mathbf{y}-\mathbf{X}b)'(\mathbf{y}-\mathbf{X}b)+\lambda 1_p'|b| = \mathrm{RSS}(b)+\lambda 1_p'|b|$$

  where $\lambda > 0$, and 1_p is a $p-$vector of ones
- Note: the Lasso penalization makes the solutions nonlinear in the $\mathbf{y}$, and there is no closed form expression as in ridge regression. An algorithm for estimation is discussed later.
- Suppose that the $\mathbf{X}'s$ are orthonormal: $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$ (this can be achieved by transforming the inputs into $\mathbf{XVD}^{-1}$). Then we have

$$\widehat{\beta}_{r,j} = \widehat{\beta}_j/(1+\lambda)$$

$$\widehat{\beta}_{l,j} = \begin{cases} 0, \text{ if } |\widehat{\beta}_j| \leq \lambda \\ \widehat{\beta}_j - \lambda \mathrm{sign}(\widehat{\beta}_j), \text{ otherwise} \end{cases}$$

  where $\widehat{\beta}_{l,j}$ is the Lasso estimator of β_j. Matters are more complicated in the general case but still Lasso sets small coefficients exactly to 0.

Lasso (cont'd)

- The Lasso enjoys the so-called oracle property, i.e. it is consistent in both parameter estimation and variable selection.
- However, no standard errors are available for the Lasso estimator. Hence, no testing is possible.
- Lasso performs well when the coefficient vector β is indeed sparse, i.e. it contains many zeros, and the $\mathbf{X}'s$ are not highly correlated.
- Ridge performs best when β contains many small coefficients, and in the presence of high correlation amongst the $\mathbf{X}'s$.
- An efficient algorithm for Lasso estimation is the LAR (Least-Angle Regression) Lasso. It provides with the entire Lasso path, i.e. the sequence of $\widehat{\beta}_{l,j}$ according to the value of λ.

LAR-Lasso

1. Center and standardize the $\mathbf{X}'s$. Start with the residual vector $\mathbf{r} = \mathbf{y} - \overline{y}1_N$ and $[\beta_1, ..., \beta_p]' = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 toward the OLS coefficient of $\mathbf{r}$ on $\mathbf{x}_j$ until some $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Update $\mathbf{r}$, and move β_j and β_k toward the OLS coefficients of $\mathbf{r}$ on $[\mathbf{x}_j, \mathbf{x}_k]$ until some $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. This is the OLS solution.

- A simple modification of LAR provides the entire Lasso path: If a non-zero coefficients hits zero, drop its variable from the active set of predictors and recompute the current joint least squares direction.

Principal Component Regression

- PCR (Principal Component Regression) consists in a OLS regression of $\mathbf{y}$ on the first $M(< p)$ principal components

$$\mathbf{Z}_M = [\mathbf{z}_1, ..., \mathbf{z}_M]$$

- This regression gives the coefficient vector

$$\widehat{\theta}_{1:M} = (\mathbf{Z}'_M \mathbf{Z}_M)^{-1} \mathbf{Z}'_M \mathbf{y} = \mathbf{D}_M^{-2} \mathbf{Z}'_M \mathbf{y} = \mathbf{D}_M^{-2} \mathbf{V}'_M \mathbf{X}' \mathbf{y}$$

where $\mathbf{D}_M^2$ is the diagonal matrix of the largest M eigenvalues and $\mathbf{Z}_M$ is the matrix of the associated eigenvectors.

- The corresponding estimate of β is obtained as

$$\widehat{\beta}_{pcr} = \mathbf{V}_M \widehat{\theta}_{1:M}$$

Partial Least Squares

- The PLS (Partial Least Squares) factors $\mathbf{F}_M = [\mathbf{f}_1, ..., \mathbf{f}_M]$ are orthogonal linear combinations of the $\mathbf{X}'s$ that maximize their covariances with the target variable $\mathbf{y}$. They are iteratively computed as follows

$$\begin{aligned}
\mathbf{U}_{(0)} &= \mathbf{X} \text{ (centered and standardized)}, \\
\mathbf{U}_{(j)} &= \mathbf{U}_{(j-1)} - \mathbf{f}_j \phi_j' = \mathbf{X} - \sum_{m=1}^{j} \mathbf{f}_m \phi_m', \; j = 1, ..., M \\
\mathbf{f}_j &= \mathbf{U}_{(j-1)} \omega_j, \\
\omega_j &= \mathbf{U}_{(j-1)}' \mathbf{y}, \\
\phi_j' &= (\mathbf{f}_j' \mathbf{f}_j)^{-1} \mathbf{f}_j' \mathbf{U}_{(j-1)}
\end{aligned}$$

- The fitted values of the PLS regression are obtained as

$$\widehat{\mathbf{y}}_{pls} = \mathbf{F}_M (\mathbf{F}_M' \mathbf{F}_M)^{-1} \mathbf{F}_M' \mathbf{y}$$

PCR vs. PLS

- The closed form solution of the PLS estimator is

$$\widehat{\beta}_{pls} = \Omega_M(\Omega_M'\mathbf{X}'\mathbf{X}\Omega_M)^{-1}\Omega_M'\mathbf{X}'\mathbf{y}$$

  where $\Omega_M = [\omega_1, ..., \omega_m]'$, which is similar to

$$\widehat{\beta}_{pcr} = \mathbf{V}_M(\mathbf{V}_M'\mathbf{X}'\mathbf{X}\mathbf{V}_M)^{-1}\mathbf{V}_M'\mathbf{X}'\mathbf{y}$$

- Differently from the principal components, the PLS factors take into account of the co-variability between the target and the predictors.
- Both PCR and PLS are consistent under the so-called Helland & Almoy condition, i.e. $\mathrm{Cov}(X, y)$ is a linear combination of M eigenvectors of $\mathrm{Var}(X)$ (not necessarily those associated with the largest eigenvalues). This implies that the eigenvalues of $\mathbf{X}'\mathbf{X}$ offer no guidance on the choice of the principal components to use in PCR.
- Whereas $\widehat{\beta}_{pcr}$ is a linear estimator, $\widehat{\beta}_{pls}$ is not. Indeed the PLS weights Ω_M depend on both $\mathbf{X}$ and $\mathbf{y}$, whereas the eigenvectors $\mathbf{V}_M$ depend on $\mathbf{X}$ only. This considerably complicates inference for PLS regression.