Inferences of Lasso Regression

In the previous video, we have used least squares to estimate inference questions. Here we are using Lasso to answer the inference question, which is,

How does the predicted value of Y change if we increase the component D of X by a unit, holding W, the other components of X fixed?

The answer is the population regression coefficient β1 corresponding to the regressor D.

A quick recap of the partialling out approach,

Write the regression equation as

$$Y = \beta_1 D + \beta_2' W + \epsilon$$

where D is the target regressor and W’s are p controls

We recall from an earlier video that after partialling-out we ended up with the simplified regression equation,

$$\tilde{Y} = \beta_1 \tilde{D} + \epsilon, \quad \mathrm{E}\epsilon\tilde{D} = 0.$$

Here Y˜ and D˜ are the residuals that are left after partialling-out the linear effect of W.

Y˜ can be calculated by predicting Y using W only and finding its residuals. i.e. removing the dependence of W on Y.

$$\tilde{Y} = Y - \gamma_{YW}' W, \quad \gamma_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} \mathrm{E}(Y - \gamma' W)^2$$

The first part of the equation says we are fitting the model with the dependent variable as Y and regressors as W and finding the residuals Y˜. Here γ'are the coefficients and the second part of the equation says we are going to find the coefficients such that the mean square error is small.

D˜ can be calculated by predicting D using W and finding its residuals, i.e. removing the dependence of W on D,

$$\tilde{D} = D - \gamma'_{DW} W, \quad \gamma_{DW} = \arg \min_{\gamma \in \mathbb{R}^p} \mathrm{E}(D - \gamma' W)^2$$

This allows us to obtain β1 as the coefficient in the linear regression of the Y˜ on D˜ in the population. We call this the partialling-out procedure.

For estimation purposes, we have a random sample of Yi's and Xi's of n observations. Our main idea here is that we will mimic in the sample the partialling-out procedure in the population. Previously, when p/n was small, we employed least squares as the prediction method in the partialling-out steps. Here p/n is not small, and we employ instead the Lasso method in the partialling-out steps.

The only difference here is for the partialling out approach while finding the residuals of Y˜ and D˜ we are going to use the lasso approach or penalize the coefficients otherwise everything is the same.

$$\widehat{\gamma}_{YW} = \arg \min_{\gamma \in \mathbb{R}^p} \sum_i (Y_i - \gamma' W_i)^2 + \lambda_1 \sum_j |\gamma_j|$$

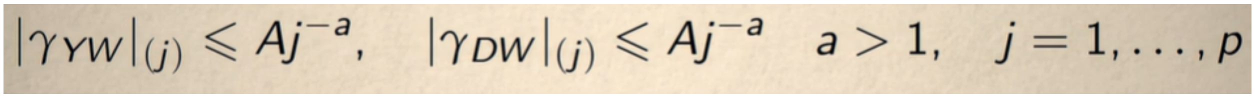
$$\widehat{\gamma}_{DW} = \arg \min_{\gamma \in \mathbb{R}^p} \sum_i (D_i - \gamma' W_i)^2 + \lambda_2 \sum_j |\gamma_j|$$

Here we can see in the equation we added a penalty term except that everything is the same.

$$\check{\beta}_1 = \arg \min_{b_1 \in \mathbb{R}} \mathbb{E}_n (\check{Y}_i - b_1 \check{D}_i)^2 = (\mathbb{E}_n \check{D}_i \check{D}_i)^{-1} \mathbb{E}_n \check{D}_i \check{Y}_i$$

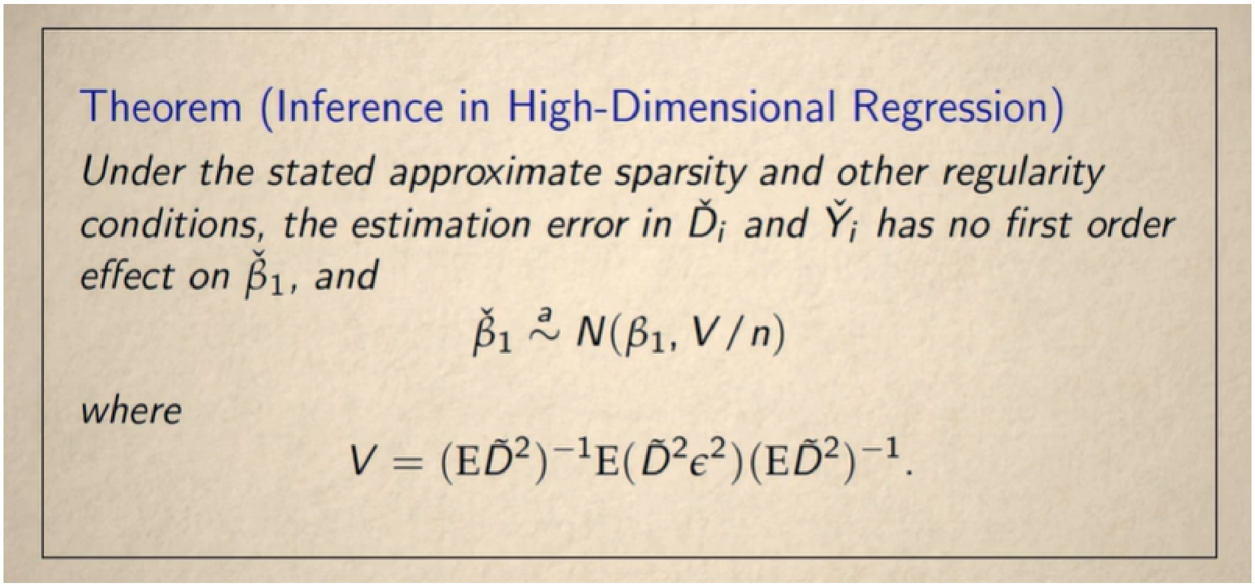
Once we get the residuals for Y˜ and D˜ we are fitting the least-squares and finding the coefficient b which is required.

Our partialling-out procedure with Lasso relies on approximate sparsity in the two partialling out steps. Indeed, theoretically, the procedure will work well if the population regression coefficients in the two partialling out steps are approximately sparse, with a sufficiently high speed of decrease of the sorted coefficients.



$$|\gamma_{YW}|_{(j)} \leqslant Aj^{-a}, \quad |\gamma_{DW}|_{(j)} \leqslant Aj^{-a} \quad a > 1, \quad j = 1, \ldots, p$$

Here the coefficients of γ_{YW} and γ_{DW} should decay fast enough.



Theorem (Inference in High-Dimensional Regression)

Under the stated approximate sparsity and other regularity conditions, the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\check{\beta}_1$, and

$$\check{\beta}_1 \overset{a}{\sim} N(\beta_1, V/n)$$

where

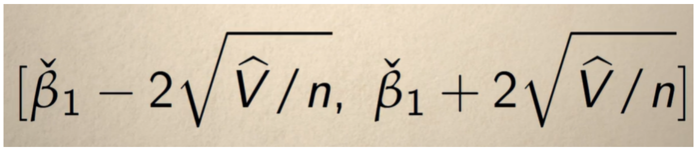
$$V = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}(\tilde{D}^2\epsilon^2)(\mathrm{E}\tilde{D}^2)^{-1}.$$

The following theoretical result states under the stated approximate sparsity and other regularity conditions, the estimation error in Dˇ i and Yˇ i has a negligible effect on βˇ 1 and βˇ 1 is approximately distributed as a Normal with mean β1 and variance V/n.

we can say that the estimator βˇ 1 concentrates in a √ V/n neighborhood of β1, with deviations controlled by the normal law.

We can define the standard error of βˇ 1 as √ Vˇ/n, where Vˇ is an estimator of V. This is our quantification of uncertainty about β1.

We then provide the approximate 95% confidence interval for β1, which is given by the estimate βˇ 1 plus/minus two standard errors,



$$[\check{\beta}_1 - 2\sqrt{\widehat{V}/n}, \ \check{\beta}_1 + 2\sqrt{\widehat{V}/n}]$$

Here we have constructed the confidence interval for β1.