Problem Set 5

Tao Wu

November 10, 2023
Professor Suleyman Taspinar
ECON 387: Advanced Econometrics

1 Question 1

(20 pts) Define the following terms in your own words.

(a) Big data gives rise to many applications including prediction with many independent variables in the model. To know which model produces the best prediction results, we need a performance metric of predictive accuracy to rely on so that we can tell. The mean squared prediction error (MSPE) is one that takes the average squared difference between the predicted values and the actual (observed) values of a dependent variable.

$$MSPE = E((Y^{oos} - \hat{Y}(X^{oos}))^2$$

The smaller the value of MSPE, the better the predictive power the model shows.

(b) Dividing the data we have into 2 subsets. One is called a training set from which we obtain model coefficients and the other one is called a validation dataset to which we apply the estimated model to get predicted out-of-sample values. The split-sample estimator of MSPE in mathematical form

$$M\hat{S}PE_{split_sample} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (Y_i^{oos} - \hat{Y}_i^{oos})^2$$

n_{test} represents the number of observations or the size of the validation dataset.

(c) The m-fold cross-validation estimator of MSPE utilizes a similar idea to the split-sample estimator. We first divide the sample into m, a finite large number, equal subsamples, and use one of them as the validation sample. Then put together (m - 1) subsamples as the training sample to estimate the model coefficients and finally test the estimated model on the validation subsample to estimators of β.

Repeat the process by choosing one of the m subsamples as a validation sample, which is different than the first validation subsample. It's like bootstrapping existing data to get more predictive bang for the data buck.

(d) To improve the accuracy of parameter estimates, a shrinkage estimator shifts the OLS model coefficients to be close to zero by reducing their variance to make up for the introduced squared bias. It's a tradeoff for the better accuracy.

(e) Ridge regression and Ordinary linear regression (OLS) are two different methods for estimating the parameters of a linear regression model. OLS gives us an unbiased estimator with a high MSPE under Gauss-Markov's assumptions. Since MSPE is a predictive metric that the minimum value wins, the Ridge regression estimator prevents estimated parameters from growing too large by introducing bias in the form of a penalty term. The purpose of ridge regression is to reduce the variance in the model at the cost of introducing some bias because we are dealing with a high-dimensional dataset that has multicollinearity and overfitting issues. Note that when $\lambda_r = 0$, the ridge regression estimator is just the OLS estimator.

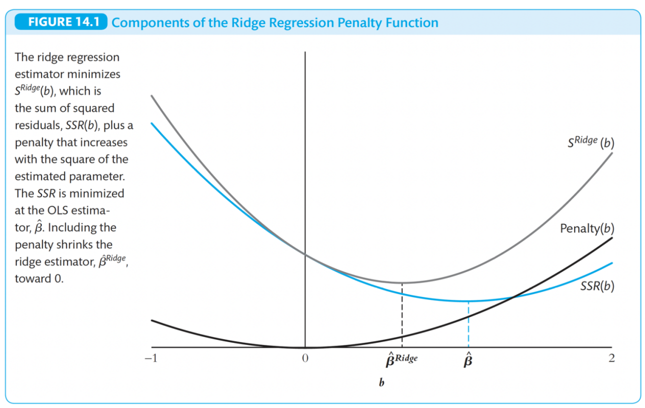
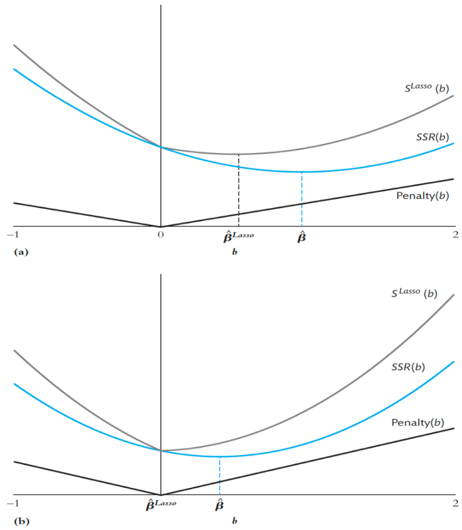


Figure 1: Ridge regression penalty function

(f) Lasso regression shares similar benefits with Ridge regression. Both methods try to accomplish the same goal which is to improve the predictive performance of a model by reducing the variance of the model parameters. However, the difference lies with Lasso reducing some of the coefficients to be exactly zero by taking the absolute values of the coefficients. Essentially, Lasso regression drops some of the independent variables from the model. Therefore, if we believe that a handful of regressors are important and provide the biggest share of predictive power, the Lassor regression estimator is a great technique to have a better prediction in terms of accuracy.

FIGURE 14.3 The Lasso Estimator Minimizes the Sum of Squared Residuals Plus a Penalty That Is Linear in the Absolute Value of *b*

For a single regressor, (a) when the OLS estimator is far from zero, the Lasso estimator shrinks it toward 0; (b) when the OLS estimator is close to 0, the Lasso estimator becomes exactly 0.



2 Question 2

(20 pts) Using data from a random sample of elementary schools, a researcher regresses average test scores on the fraction of students who qualify for reduced-price meals. The regression indicates a negative coefficient that is highly statistically significant and yields a high adjusted R2. Is this regression useful for determining the causal effect of school meals on student test scores? Why or why not? Is this regression useful for predicting test scores? Why or why not?

This is **not** a regression useful for determining the causal effect of school meals on student test scores despite the highly statistical significance and high adjusted R^2. Without controlling for any relevant variables, especially when there is available data on other regressors, claiming a causal effect is a farce. Even though control variables are present in the model, it is a tall bar to assume that the fraction of students who qualify for reduced-price meals does not correlate with any other unobserved factors in the error term making its coefficient an unbiased estimator. However, when it comes to the business of prediction, I think this regression might be useful for predicting test scores because accurate prediction trumps attributing causality. Highly statistical significance and high adjusted R^2 tell us that the regressor has a high degree of predictive power on students' test scores.

3 Question 3

(20 pts) Ridge and Lasso are two regression estimators based on penalization. Explain how they are similar and how they differ.

The goal of prediction is to always have a good out-of-sample fit, which means that we want the lowest possible MSPE metric. The Ridge regression introduces bias in the form of a penalty term which shrinks the ridge regression estimator toward 0, but not at 0, by reducing the variance in the model at the cost of introducing some bias because we are dealing with a high-dimensional dataset that has multicollinearity and overfitting issues.

Lasso regression shares similar benefits with Ridge regression. Both methods try to accomplish the same goal which is to improve the predictive performance of a model by reducing the variance of the model parameters. However, the difference lies with Lasso reducing some of the coefficients to be exactly zero by taking the absolute values of the coefficients. Essentially, Lasso regression drops some of the "useless" independent variables in a sparse model. Therefore, if we believe that a handful of regressors are important enough and provide the biggest share of predictive power, the Lassor regression estimator is a great technique to have a better prediction in terms of accuracy.

4 Question 4

(a) Computing the predicted value for an out-of-sample observation requires standardizing the observation using the in-sample mean and variance of each predictor.

Out-of-Sample predictions using principal components regression

- The predicted value of $Y^{\star v}$ is computed as follows:

1. Compute the principal components in the estimation (training) sample:
    1. Compute the demeaned Y and standardized X for the in-sample observations on $Y^\star$ and $X^\star$'s.
    2. Compute the in-sample principal components of X's: $PC_1, PC_2, \ldots, PC_{\min(k,n)}$.
2. Given p, estimate the regression coefficients in equation (4), call these estimates $\widehat{\gamma}_1^{pc}, \widehat{\gamma}_1^{pc}, \ldots, \widehat{\gamma}_p^{pc}$.
3. Compute the out-of-sample values of the principal components:
    1. Standardize the out-of-sample predictors $X^{\star v}$'s using the in-sample mean and standard deviation from step 1.1. Denote this transformed observation as X^v's.
    2. Compute the principal components for the out-of-sample observation using the in-sample weights from step 1.2: $PC_1^v, PC_2^v, \ldots, PC_p^v$.
4. Compute the predicted value for the out-of-sample observation as $\widehat{Y}^{\star v} = \bar{Y}^\star + \sum_{j=1}^{p} \widehat{\gamma}_j^{pc} PC_j^v$, where $\bar{Y}^\star$ in-sample mean of $Y^\star$ from step 1.1.

i. Let's compute X^{oos} values transformed (standardized):

$$RPM = \frac{0.52 - 0.6}{0.28} \approx -0.2857$$

$$TExp = \frac{11.1 - 13.2}{3.8} \approx -0.5527$$

ii. The predicted value of average test scores:

$$9.1051(\hat{Y}^{*v}) + 750.1(\bar{Y}^{*}) = 759.2051$$

(b) The actual average test score for the school is 775.3. The error for your prediction is computed as $|759.20 - 775.3| = 16.9$

(c) Suppose the regression had been estimated using the raw data for TestScore, RMP, and TExp. The value of the raw data regression intercept would be the sample mean of 750.1 for the outcome variable (TestScore). The slope coefficients for RPM would be -13.64 and 33.06 for TExp respectively.

(d)

$$\widehat{TestScore} = 750.1 + (-13.64) * 0.52 + 33.06 * 11.1 = 1109.97$$

Are predicted average test scores in a) and the raw data regression model prediction supposed to be the same?