Regression, Prediction and Shrinkage

In this assignment, we will demonstrate the comparison of prediction accuarcy among several linear predictors, i.e. ordinary least square (OLS) predictors, pre-shrunk predictors as in Copas(1983) and shrinkage predictor by cross-validation.

Simulation Setup

To demonstrate the required statements about prediction mean square error (PMSE) of the three predictors. We choose the simulation parameters as shown in the table below.

Simulation Parameter	Symbol	Value
predictor dimension	p	50
number of overall samples	n	100
number of training samples	n_{CS}	80
number of test samples	n_{VS}	20
true intercept	α	0.1
true coefficients	$\boldsymbol{\beta}$	$\in$[0.1, 5]
noise-signal-ratio	δ^2	0.01
noise variance	σ^2	343.4
simulation runs	N	100

In order to show the strength of Copas(1983) method, we would like a relatively large p with a relatively small n. Also, the error variance σ^2 should be comparable to $\boldsymbol{\beta}^{\boldsymbol{T}} V \boldsymbol{\beta}$ with non-zero noise-signal-ratio δ^2. So, as suggested by the Breiman&Friedman(1997) paper design, we set our simulation parameters as above. The true linear model parameters, $\alpha, \boldsymbol{\beta}$,are not critical, so we just made some simple choices. Moreover, for strong hypothesis testing power, we have fairly large repetition runs, N, of simulations.

As for the data statistics, for each run, the predictor variables are generated according to a normal distribution with zero mean and covariance matrix $\boldsymbol{\Xi}$,

$$\begin{aligned} &\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Xi}) \\ &\boldsymbol{\Xi} = [V_{ij}]_{p\times p}, \quad V_{ij} = r^{|i-j|}, \quad r \sim \mathcal{U}[0,1] \end{aligned}$$

Also, the model matrix $\boldsymbol{X}$ is column centered for simplicity (without loss of generality), and we have $\boldsymbol{V} = \frac{1}{n}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$. Meanwhile, for both *construction sample*(CS) and *validation sample*(VS), the data sample vector $\boldsymbol{y}$ are generated independently w.r.t each element,

$$\begin{aligned} y &= \alpha + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x} + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

So, we can see that the randomness in my simulation comes from three sources: first, the randomness between each run (rep. of dim=N); second, the randomness in the data noise (ϵ of dim=n); and finally, the randomness in the predictive factors (x of dim=p).

Additionally, to verify the correctness of my implementation, we would like to check if our estimates, $\hat{\sigma}^2$, are unbiased according to a one-sample t-test with target $\mu_0 = \sigma^2$.

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{\nu}\sum_{i=1}^{n_{CS}} (y_i - \hat{y}_i^{OLS})^2 \\ \nu &= n_{CS} - (p+1) \end{aligned}$$

And the two-sided t-test results in a p-value of 0.4777, which confirms that our estimates are unbiased.

Moreover, we also want to check if the $\boldsymbol{V}$ (as defined in Copas(1983)) is close to the true covariance matrix $\boldsymbol{\Xi}$ in terms of the Frobenius norm.

$$||\boldsymbol{V}||_F = \sum_{i=1}^{p}\sum_{j=1}^{p}|V_{ij}|^2$$
$$||\boldsymbol{\Xi}||_F = \sum_{i=1}^{p}\sum_{j=1}^{p}|\Xi_{ij}|^2$$

We conduct a two-sample t-test on the N samples of matrix norms and get the p-value of 0.8705, which confirms that we can trust the $\boldsymbol{V}$ approximation.

Now, we are safe to proceed and check the hypothesis in the statements.

Statement 1: Copas vs. OLS

We want to evaluate

H0: the PMSE of Copas pre-shrunk predictor is the same as that of OLS

HA: the PMSE of Copas pre-shrunk predictor is lower than that of OLS

We have the predictors with OLS parameter estimates, $\hat{\alpha}, \hat{\boldsymbol{\beta}}$, for CS data samples,

$$\hat{y}^{OLS} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x}$$
$$\hat{y}^{Copas} = \hat{\alpha} + \hat{K}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x}, \quad \hat{K} = 1 - \frac{(p-2)\hat{\sigma}^2\nu}{n_{CS}(\nu+2)\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{V}\hat{\boldsymbol{\beta}}}$$

and then conduct a one-sided two-sample t-test on the PMSE values obtained from N runs on the VS samples,

$$PMSE = \frac{1}{n_{VS}}\sum_{i=1}^{n_{VS}}(y_i - \hat{y}_i)^2$$

the test gives a p-value of 0.0043 which rejects the null hypothesis and shows that PMSE of Copas pre-shrunk predictor is lower than that of OLS.

Statement 2: Copas vs. Cross-validation

We want to evaluate

H0: the PMSE of Copas pre-shrunk predictor is the same as that of shrinkage estimated from cross-validation

HA: the PMSE of Copas pre-shrunk predictor is lower than that of shrinkage estimated from cross-validation

We have the predictors with OLS parameter estimates, $\hat{\alpha}, \hat{\boldsymbol{\beta}}$, for CS data samples,

$$\hat{y}^{Copas} = \hat{\alpha} + \hat{K}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x}, \quad \hat{K} = 1 - \frac{(p-2)\hat{\sigma}^2\nu}{n_{CS}(\nu+2)\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{V}\hat{\boldsymbol{\beta}}}$$
$$\hat{y}^{CV} = \hat{\alpha} + \tilde{K}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x}, \quad \tilde{K} = \frac{1}{n_{CS}}\sum_{i=1}^{n_{CS}}K_i^{CV}$$
$$K_i^{CV} = \frac{y_i}{\hat{y}_i}, \quad \hat{y}_i = \hat{\alpha}^{/i} + (\hat{\boldsymbol{\beta}}^{/i})^T\boldsymbol{x}_i$$

and then conduct a one-sided two-sample t-test on the PMSE values obtained from N runs on the VS samples, the test gives a p-value of 4.458×10^{-12} which rejects the null hypothesis and shows that PMSE of Copas pre-shrunk predictor is lower than that of shrinkage estimated from cross-validation. (Here, we set the δ^2=10^{-4} for the simulation setup.)

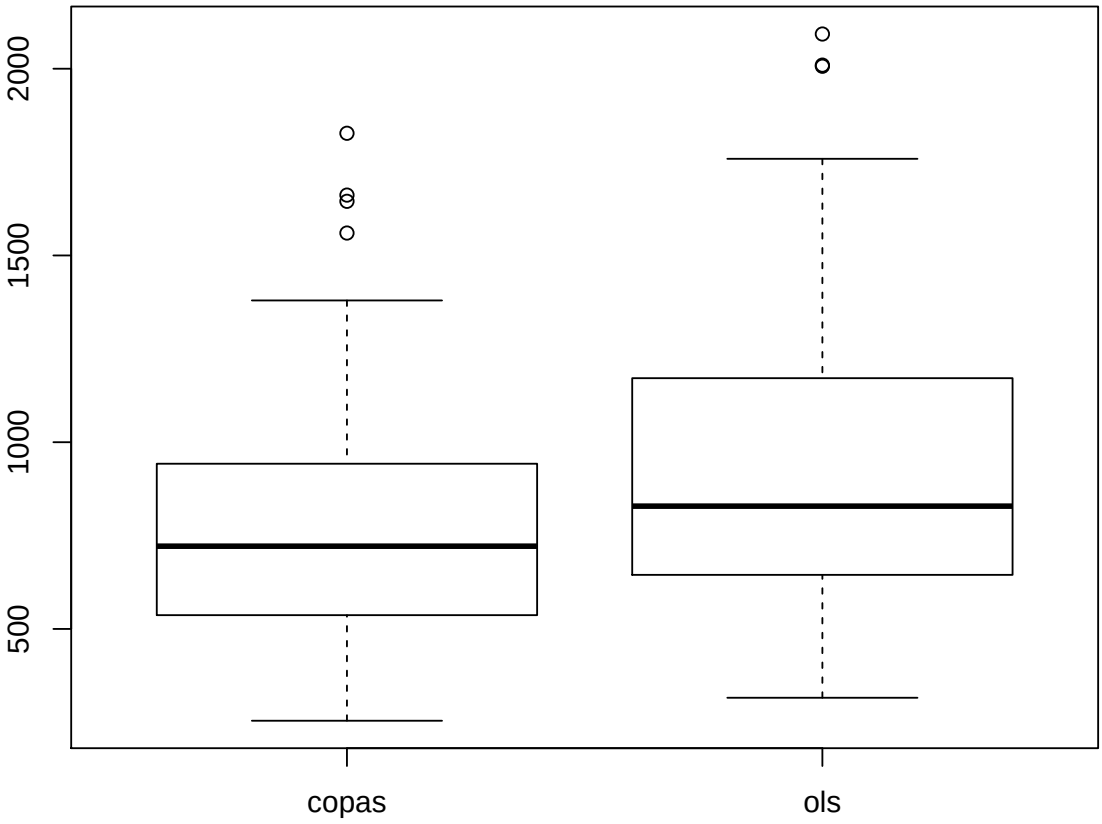

Figure 1: PMSE Copas vs. OLS

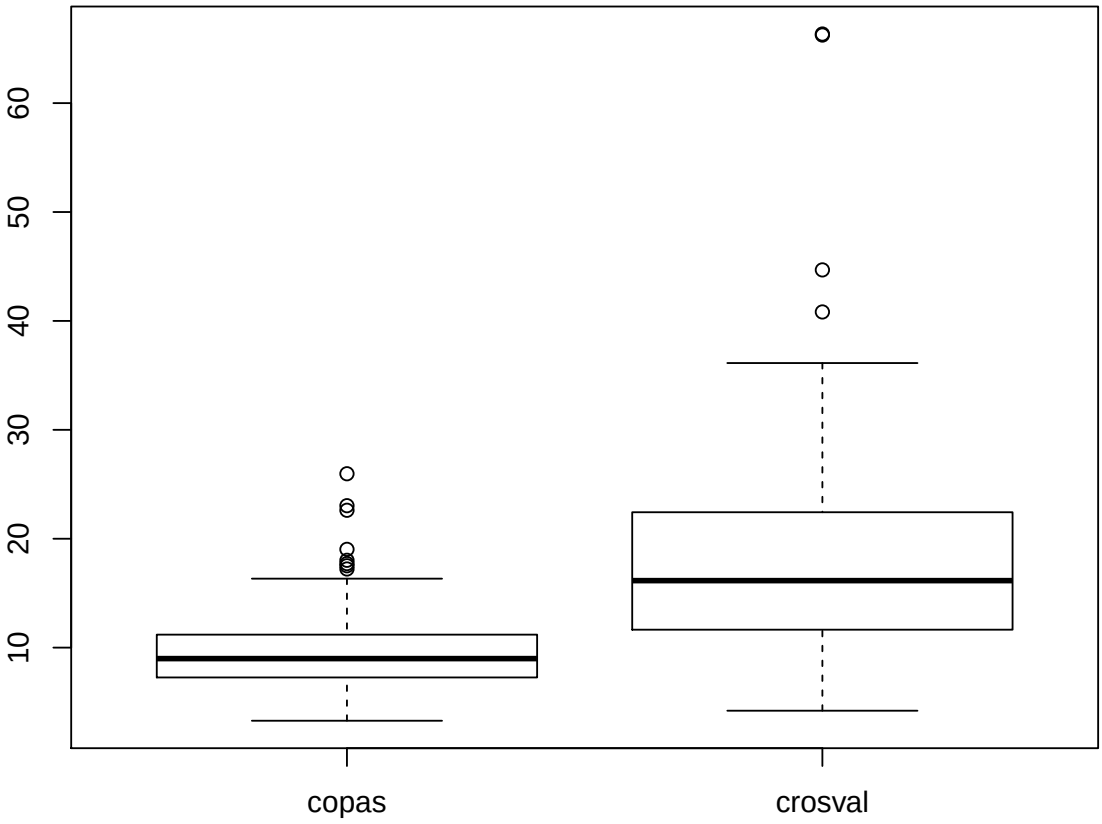

Figure 2: PMSE Copas vs. Cross-validation

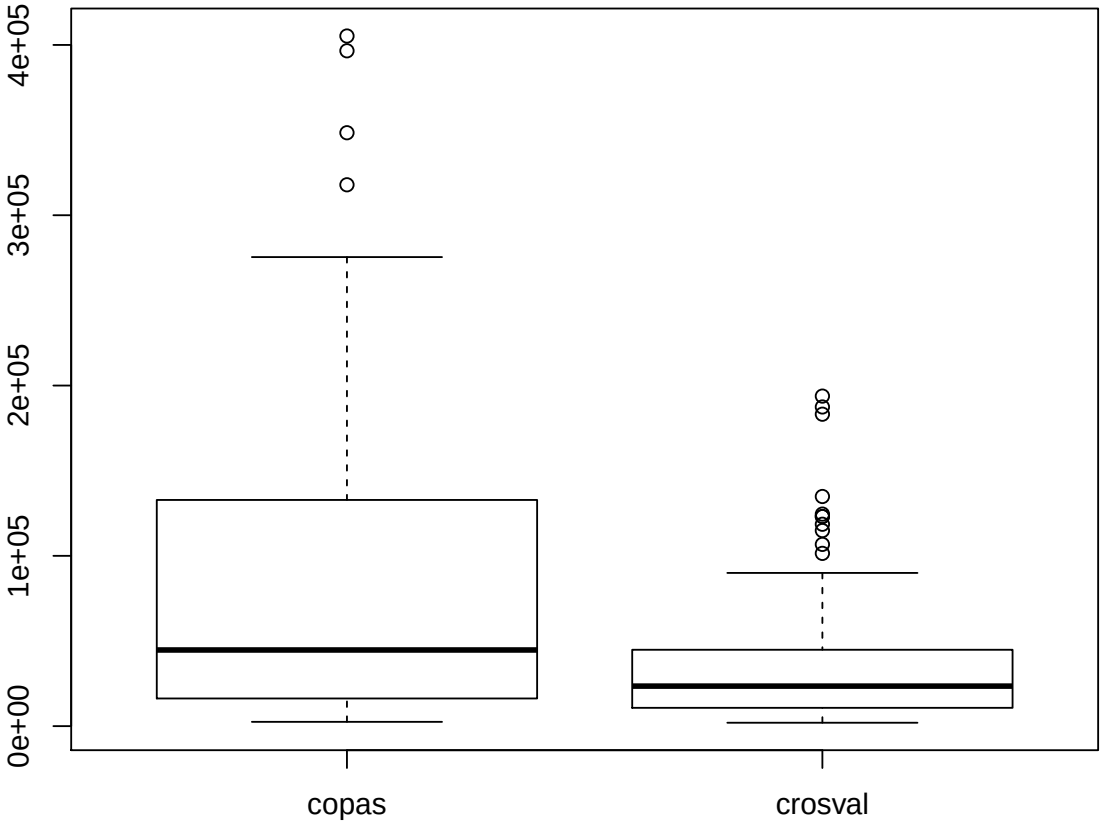

Figure 3: PMSE Copas vs. Cross-validation given non-normal noise

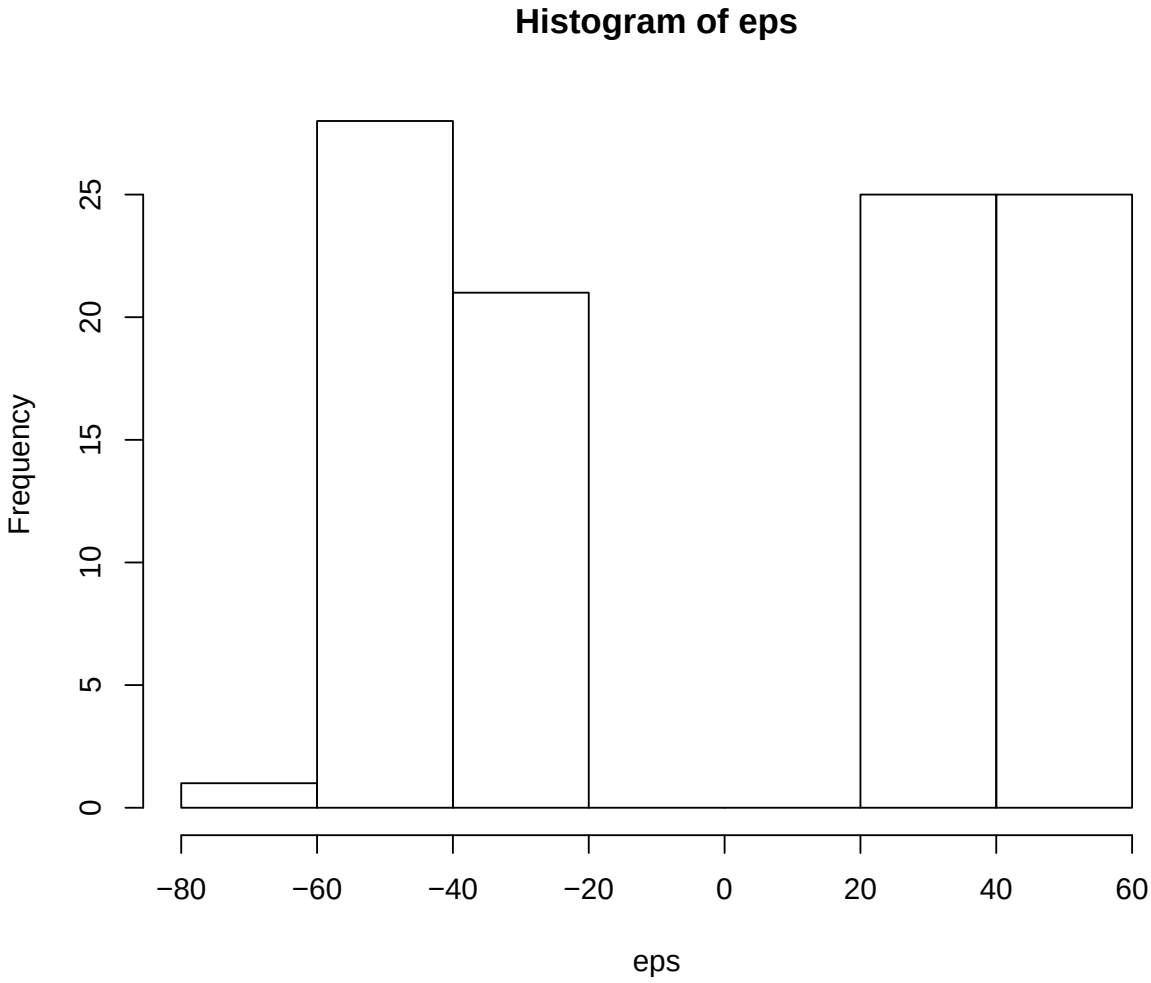

Figure 4: Histogram of eps

Statement 3: Copas vs. Cross-validation | non-normal noise

We want to evaluate

H0: the PMSE of Copas pre-shrunk predictor is the same as that of shrinkage estimated from cross-validation under non-normal noise

HA: the PMSE of Copas pre-shrunk predictor is larger than that of shrinkage estimated from cross-validation under non-normal noise

We have the same predictors as in statement 2, but here we do not have a normal noise ϵ, but a histogram shown in the Figure. Also, we need a larger δ^2 to strengthen the test, i.e. 0.04 Then we conduct a one-sided two-sample t-test on the PMSE values obtained from N runs on the VS samples, the test gives a p-value of 4.3421×10^{-6} which rejects the null hypothesis and shows that PMSE of Copas pre-shrunk predictor is larger than that of shrinkage estimated from cross-validation.