1 Mussels' muscles

The following study came from an ecological study of horsemussels sampled from the Marlborough Sounds off the coast of New Zealand, an example that is presented in section 4.4.4. from [1]. In this case, the response variable y is the logarithm of the mussel's muscle mass, and four predictors were considered; these are the logarithms of the mussel's shell height, shell width, shell length (each in millimeters), and shell mass (in grams). The sample size is $n = 79$ and $p = 4$. First, we can see that for this dataset, the cross-validation results indicate that $u = 1$ component is the appropriate one for the model (see Figure 1). In Figure 2 we can see the linear behavior for the true response vector y versus the fitted values where PLS model with $u = 1$ component was used. Next, we use a leave-one-out method to construct the prediction intervals for each observation following the next result

Corollary 1. *Under the hypothesis of Theorem 1 an approximate* $1 - \alpha$ *level prediction interval is given by*

$$PI_\alpha = \left[\widehat{\boldsymbol{\beta}}_{pls}^T \boldsymbol{X}_N \pm z_{\alpha/2} \left(V + \tau^2\right)^{1/2}\right].$$

The n prediction intervals are shown in Figure 3.

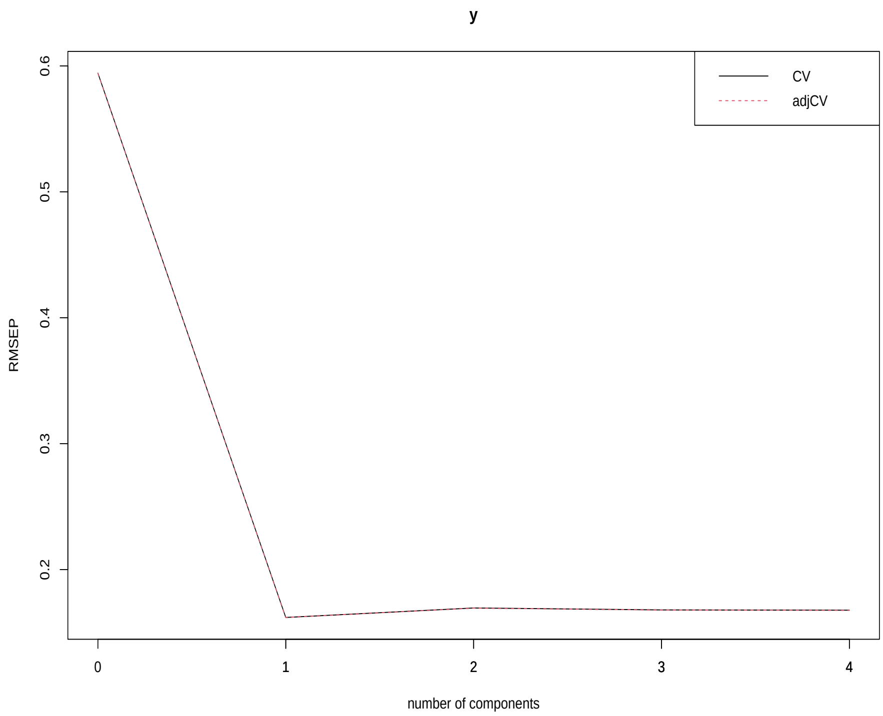

Figure 1: CV for the number of components.

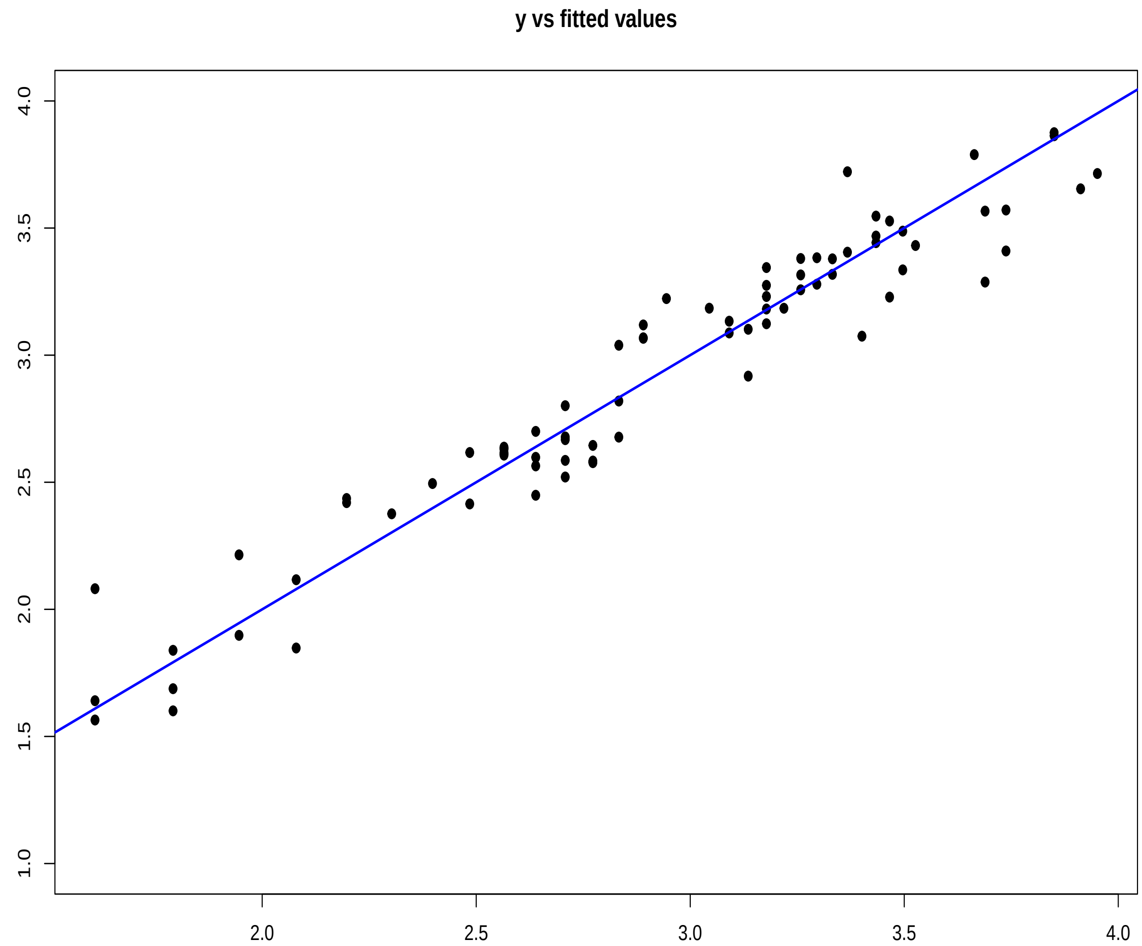

Figure 2: Plot for the response vector versus the fitted values for the PLS model $u = 1$ along with the blue line that corresponds to $y = x$.

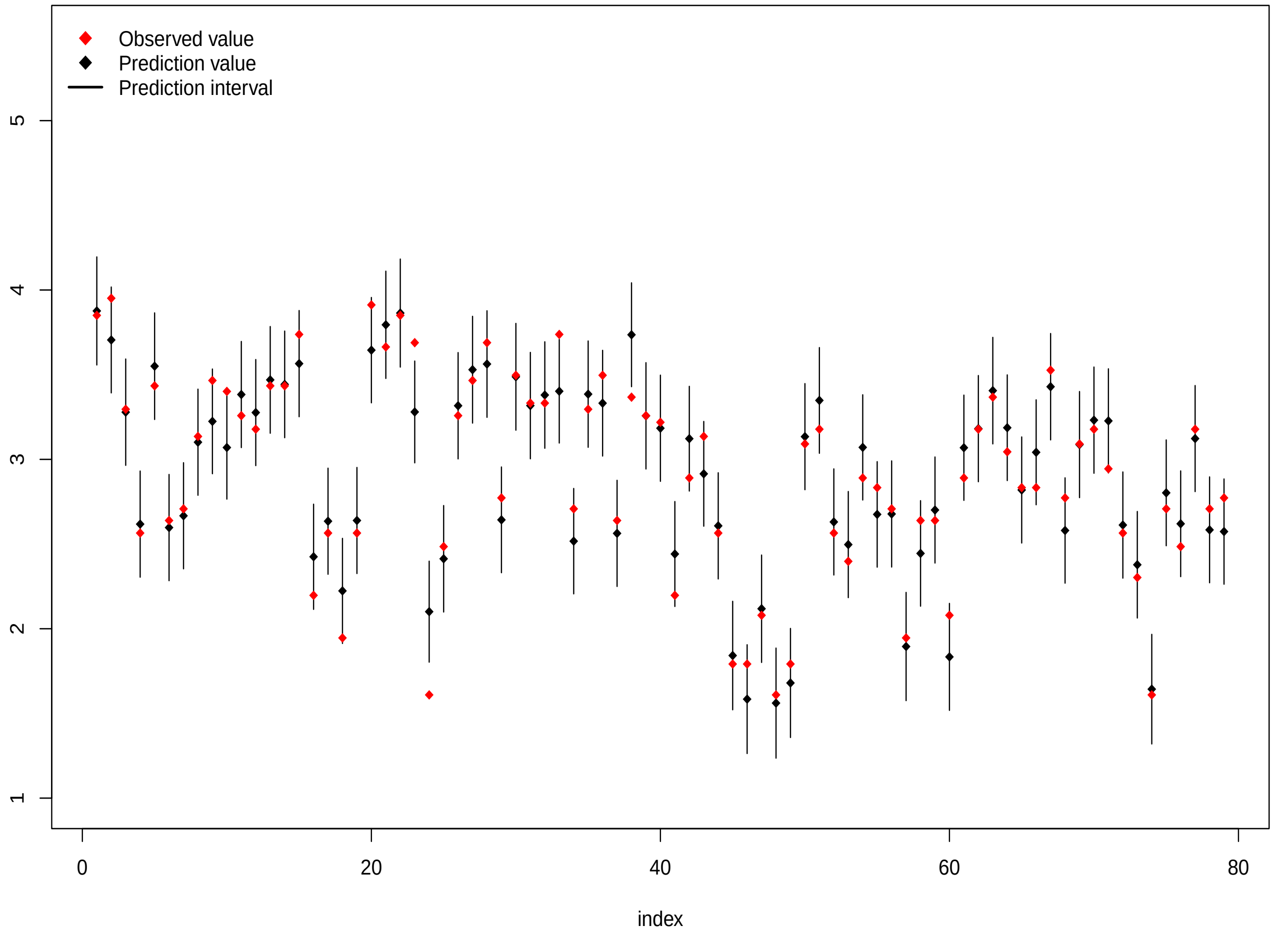

Figure 3: Intervals for the $n = 79$ observations in the mussels data. 93.67% of the intervals contain the observed value.

Now we are going to compare the classic OLS estimation with PLS estimation. Furthermore, we also add the envelope estimation with $q = 1$ obtained in section 4.4.4. from [1]

Table 1: Results for the OLS regression

	Value	Std	Confidence Interval
$\widehat{\beta}_1$	0.741	0.410	$(-0.0759, 1.559)$
$\widehat{\beta}_2$	-0.113	0.399	$(-0.909, 0.683)$
$\widehat{\beta}_3$	0.567	0.118	$(0.330, 0.803)$
$\widehat{\beta}_4$	0.170	0.304	$(-0.435, 0.776)$

Table 2: Results for the PLS regression with $u = 1$

	Value	Std	Confidence Interval
$\widehat{\beta}_1$	0.142	0.0063	$(0.129, 0.154)$
$\widehat{\beta}_2$	0.153	0.0067	$(0.140, 0.166)$
$\widehat{\beta}_3$	0.624	0.0199	$(0.585, 0.663)$
$\widehat{\beta}_4$	0.206	0.0086	$(0.189, 0.223)$

Table 3: Results for the envelope model, $q = 1$

	Value	Std	Confidence Interval
$\widehat{\beta}_1$	0.141	0.0052	$(0.130, 0.151)$
$\widehat{\beta}_2$	0.154	0.0056	$(0.143, 0.164)$
$\widehat{\beta}_3$	0.625	0.0194	$(0.587, 0.663)$
$\widehat{\beta}_4$	0.206	0.0073	$(0.191, 0.220)$

We present here the four confidence intervals for $\widehat{\beta}_i$ for $i = 1, 2, 3, 4$ with these methods. Confidence intervals with PLS method were constructed following Corollary 2

$$\left[\widehat{\boldsymbol{\beta}}_{pls}^{T}\boldsymbol{e}_i \pm z_{\alpha/2}\widehat{V}_{\boldsymbol{e}_i}^{1/2}\right]$$

where $\boldsymbol{e}_i$ is the vector with a 1 in the i-th position and 0's elsewhere.

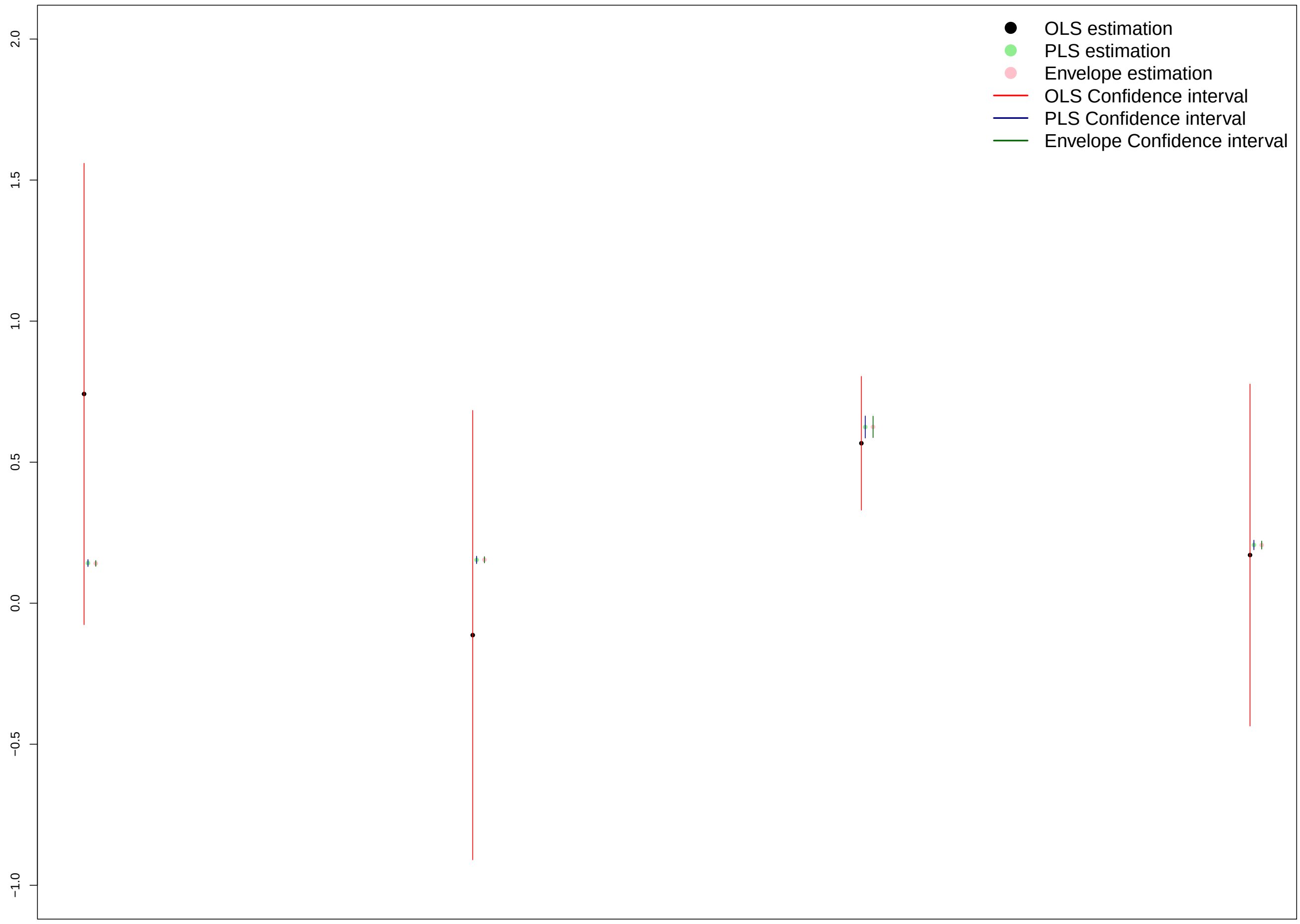

Figure 4: Comparison between the OLS, PLS and envelope estimations for $\widehat{\beta}_i$ for $i = 1, 2, 3, 4$. Observe that while the intervals for PLS and envelopes are quite similar, the intervals for OLS are significantly greater.

References

[1] Cook, R. D., *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*, Wiley Series in Probability and Statistics, ISBN = 9781119422969