A Comparison on multi-response multivariate estimation methods

Raju Rimal, Trygve Almøy and Solve Sæbø

25 Apr, 2018

Introduction

- “Big Data” is becoming a focal discussion in most of the discipline
- Massive explosion of data with informations integrated in many variables and features
- New methods and algoriths are being devised inorder to extract such information and study the relationship between different variables
- Modern inter-disciplinary research fields such as chemometrics, echonometrics and bioinformatics are handling multi-response models extensively
- This paper attempts to compare some of such methods and their performance on linear model data with specifically designed properties

Background

- Discuss some previous study on comparison specifically on multi-response setting
- Discuss the experimenal design settings on those papers
- What is new thing about this paper that other have not done

Objective

- Demonstrate a systematic comparison study using SimrelM
- Compare new estimation methods with conventional methods using data with properties particularly constructed for comparison

---

Statistical Model

- Simulation model

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right)$$

- Define transformation as $\mathbf{z} = \mathbf{R}\mathbf{x}$ and $\mathbf{w} = \mathbf{Q}\mathbf{y}$
- Equivalent latent model will be,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz} \\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz} \end{bmatrix} \right)$$

- *How much should I discuss about simrel-M??*

Exerimental Design

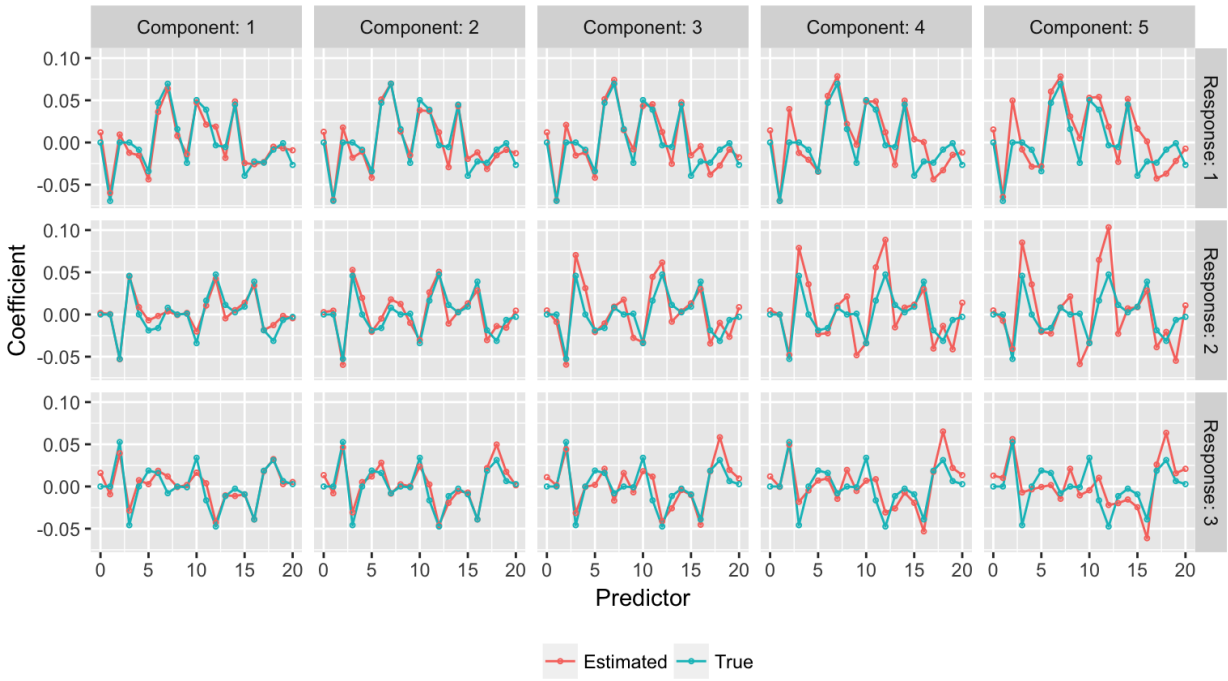
- Parameters with single level:
  - Number of observations (`n`): 100
  - Number of response variables (`m`): 3
  - Number of informative response components: 2
  - Position of predictor components relevant for response components (`relpos`): 1, 4; 2, 3
  - *Something smart* (`ypos`): 1; 2, 3
- Parameters with multiple level:
  - Number of predictor variables (`p`): 2 levels (20, 250)
  - Decay factor of eigenvalues corresponding to predictors (`gamma`): 2 levels (0.2, 0.9)
  - Decay factor of eigenvalues corresponding to response (`eta`): 2 levels (0.1, 0.8)
  - Coefficient of determination corresponding to each informative response compnents `R2` : 2 levels (0.8, 0.8; 0.4, 0.8)

Estimation Methods

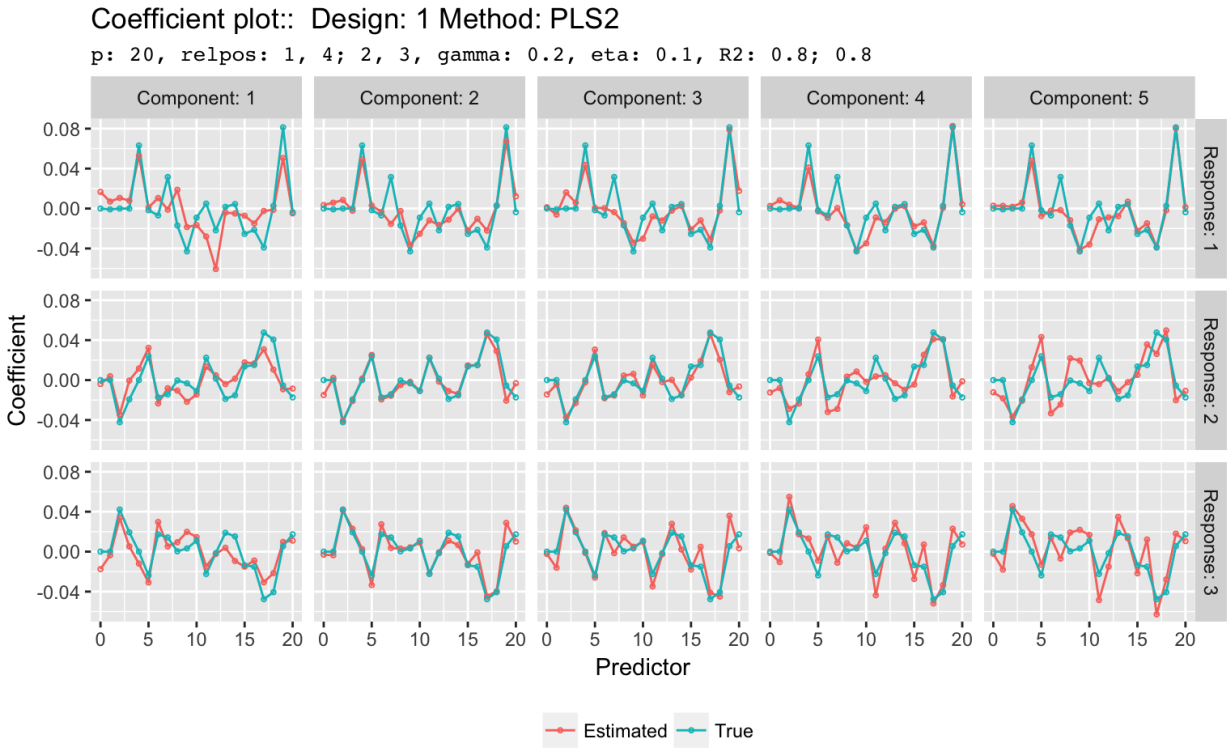
- Methods used in the study and their short description (how they estimate, what are they based on)

1. Principal Component Regression (PCR)

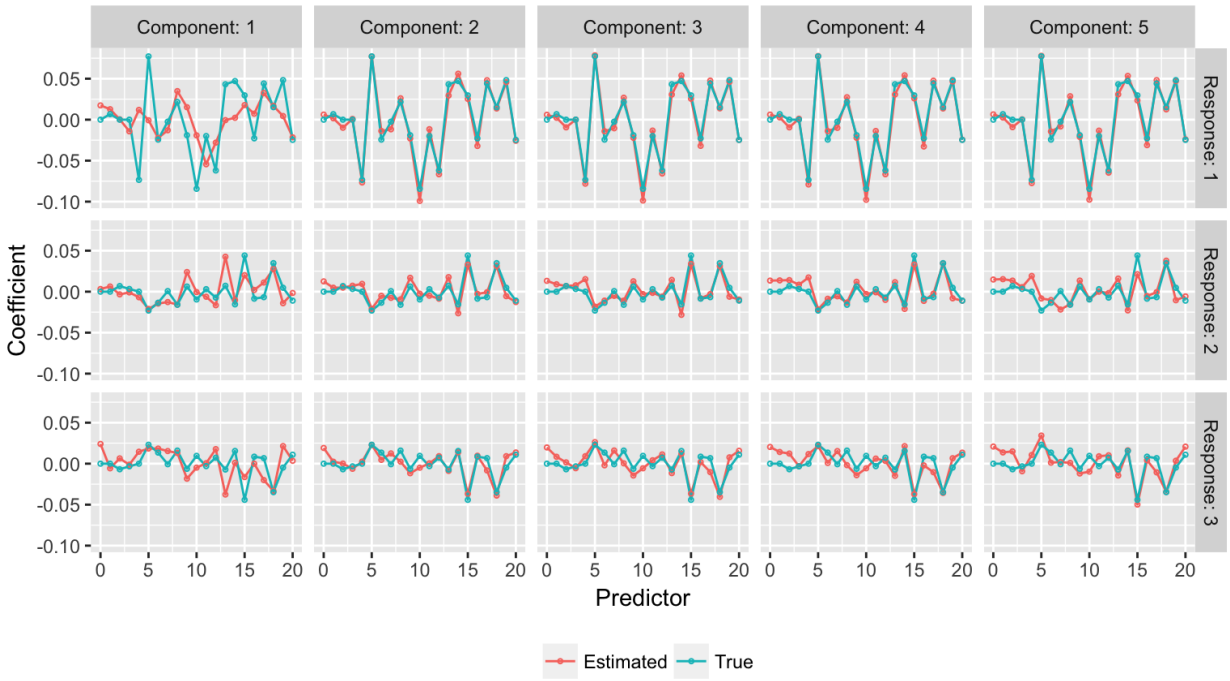
2. Partial Least Squares 1 (PLS1)



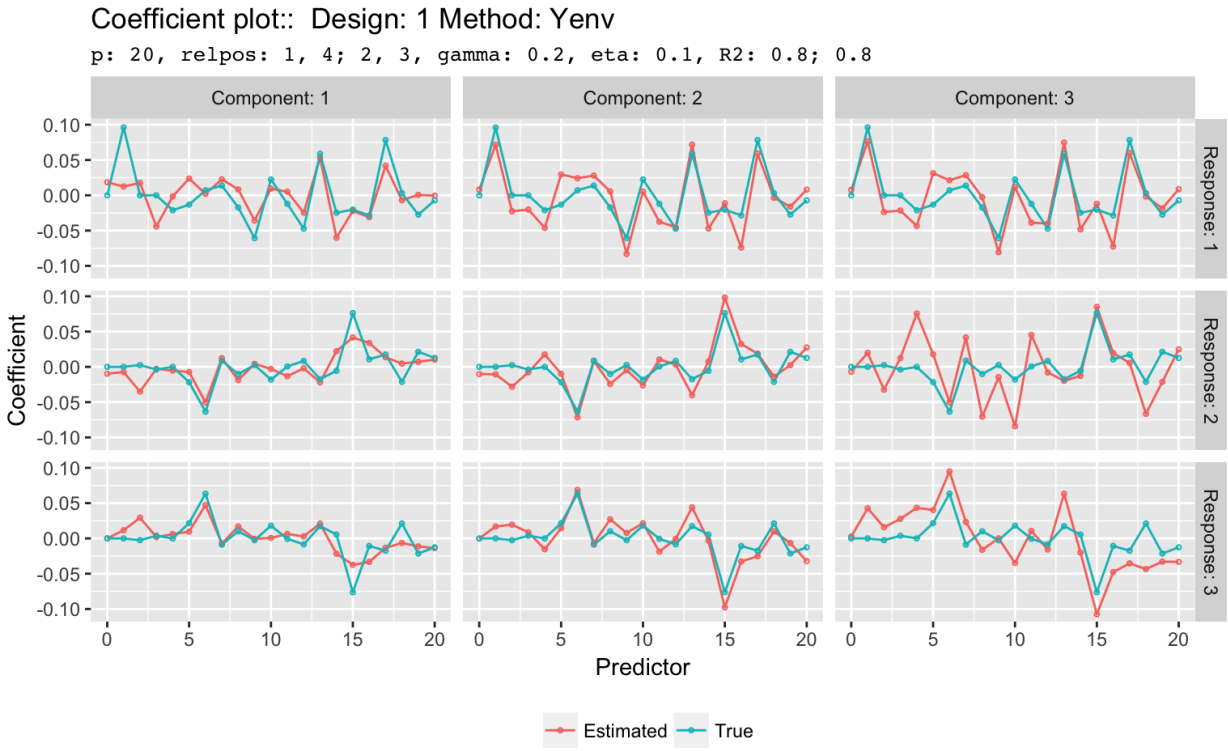
3. Partial Least Squares 2 (PLS2)



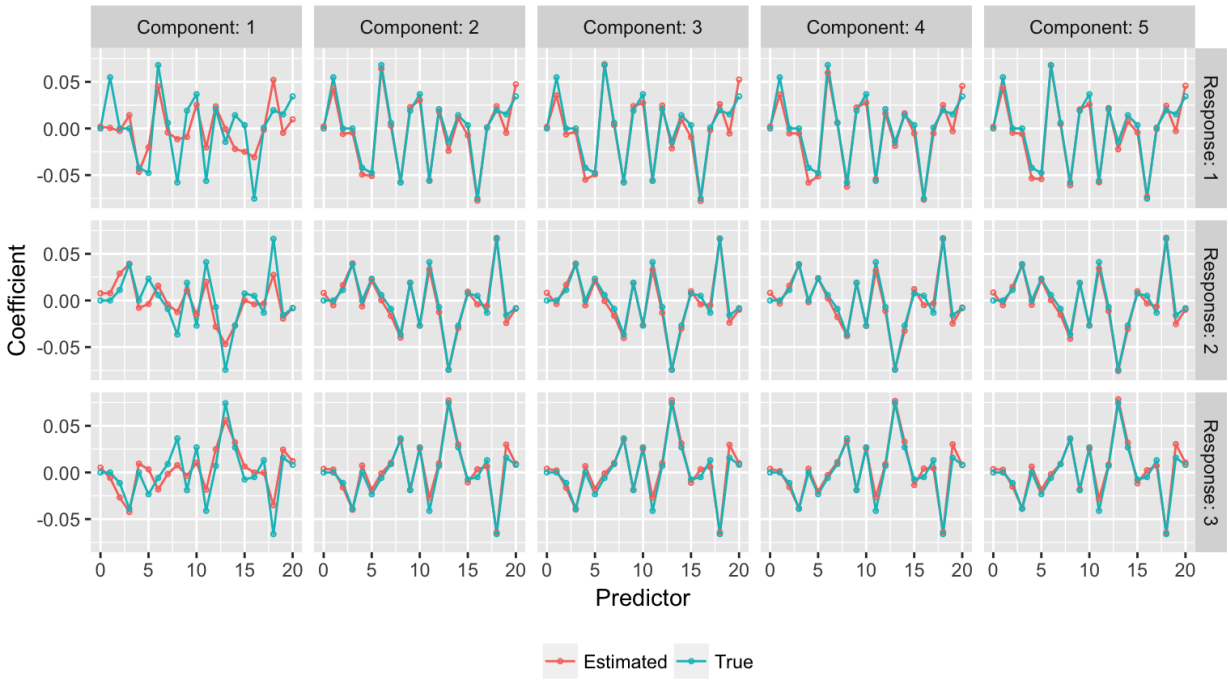
4. Envelope Estimation in Predictor Space (Xenv)



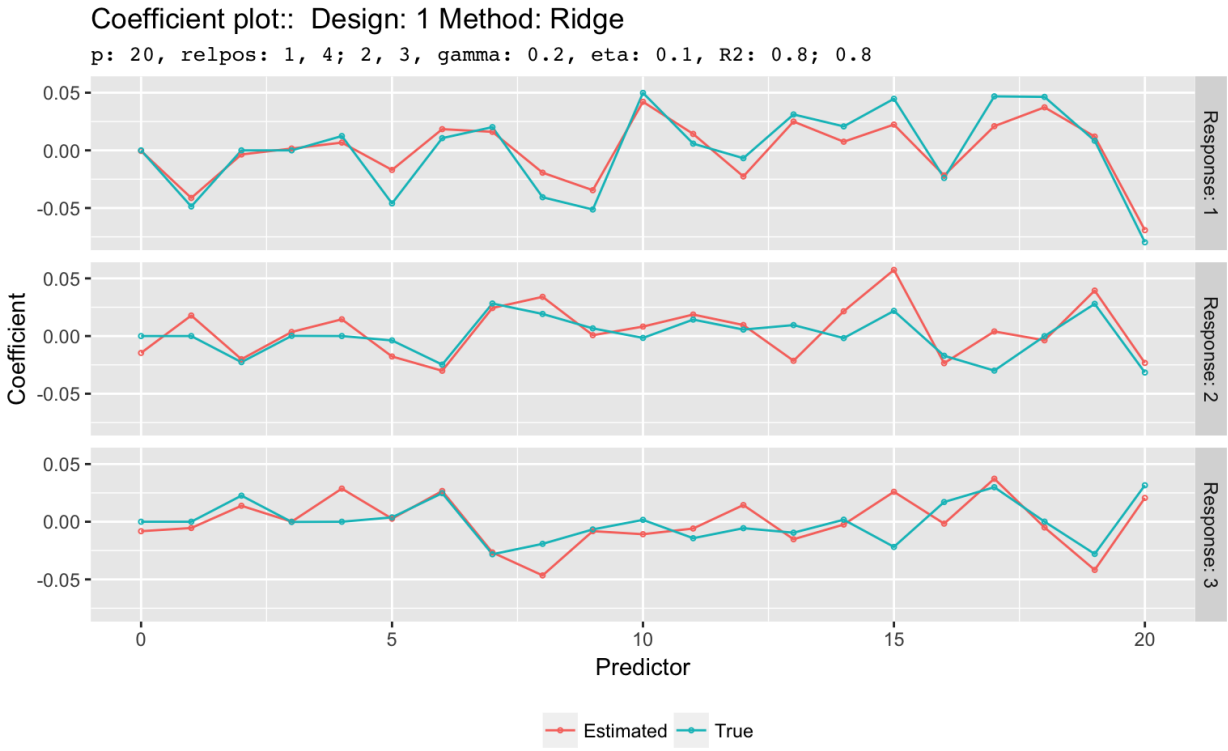
5. Envelope Estimation in Response Space (Yenv)



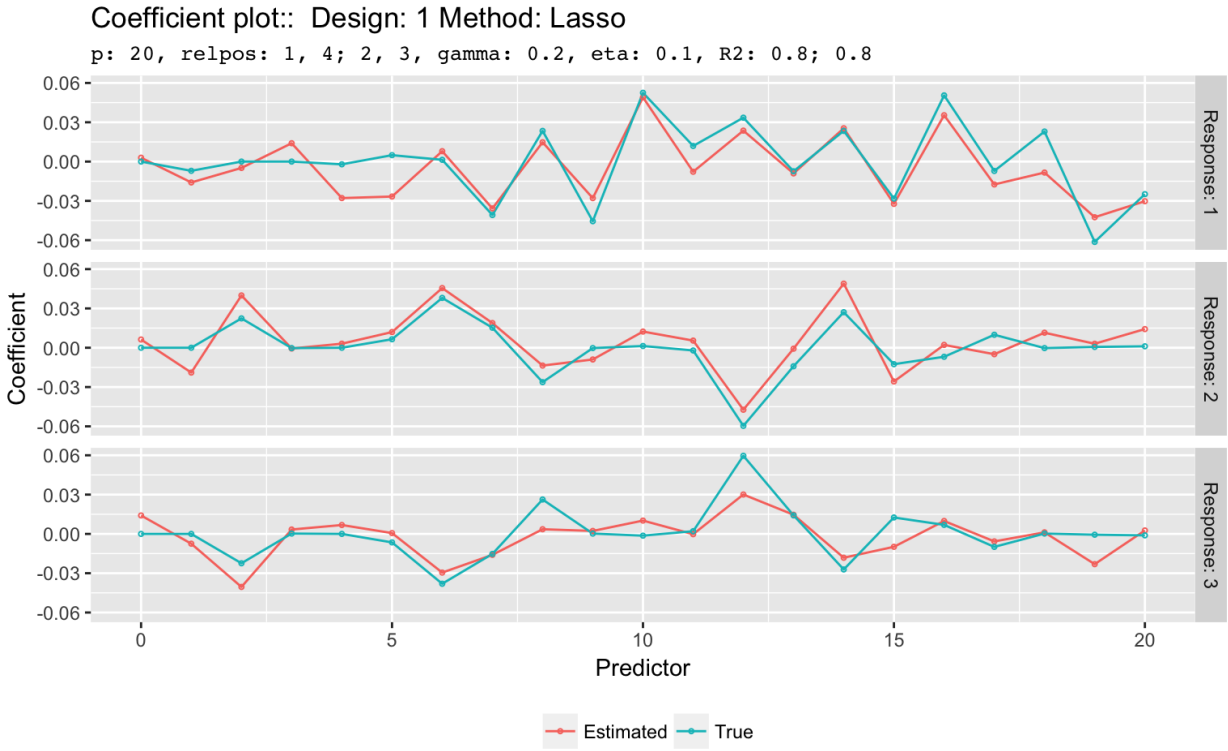
6. Simulteneous envelope estimaion (Senv)



7. Ridge Regression (Ridge)



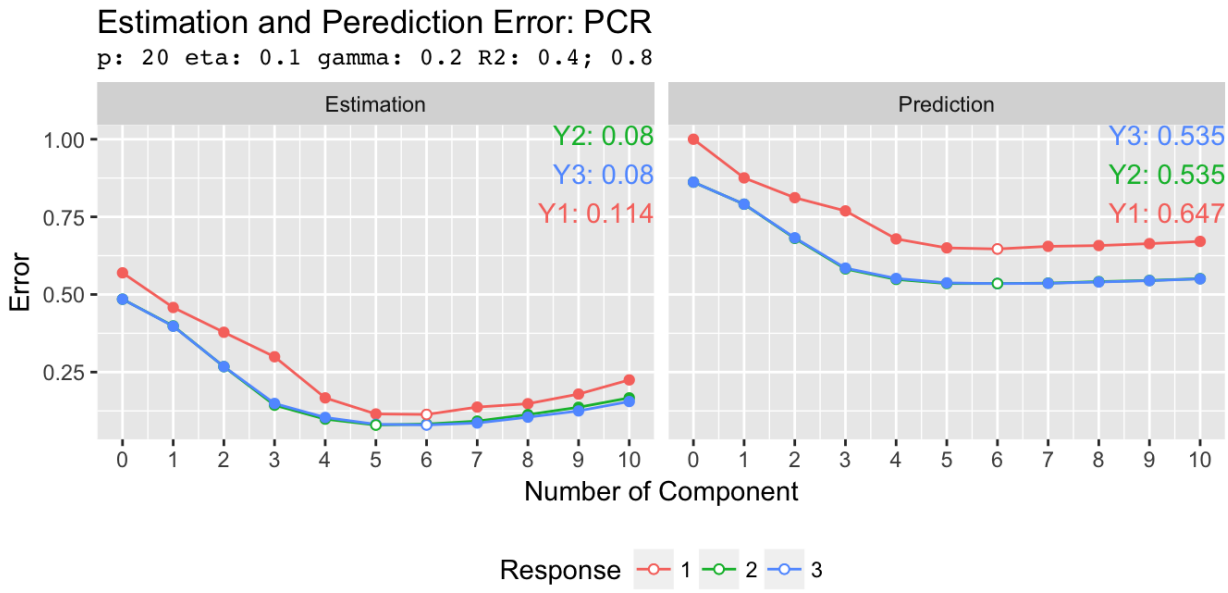
8. Lasso Regression (Lasso)

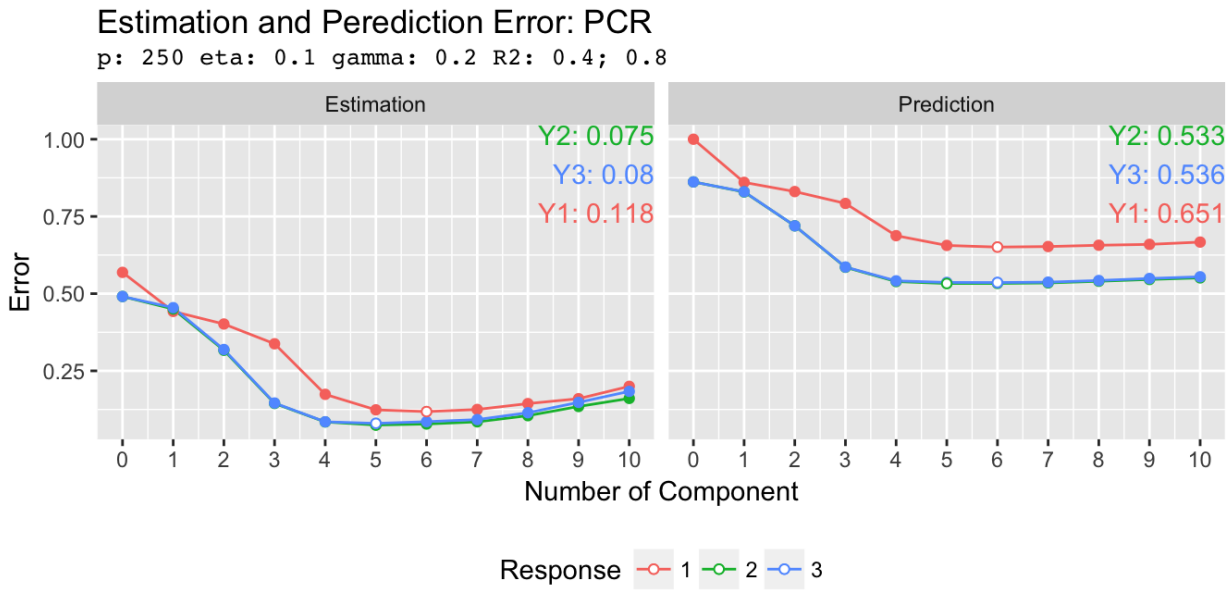


- ☐ *How details should I discuss about these methods in terms their way of estimation and difference between them*
- ☐ As *Xenv*, *Yenv* and *Senv* are based on maximum likelihood estimation, principal components of predictors explaining 99.5% of their variation are used.

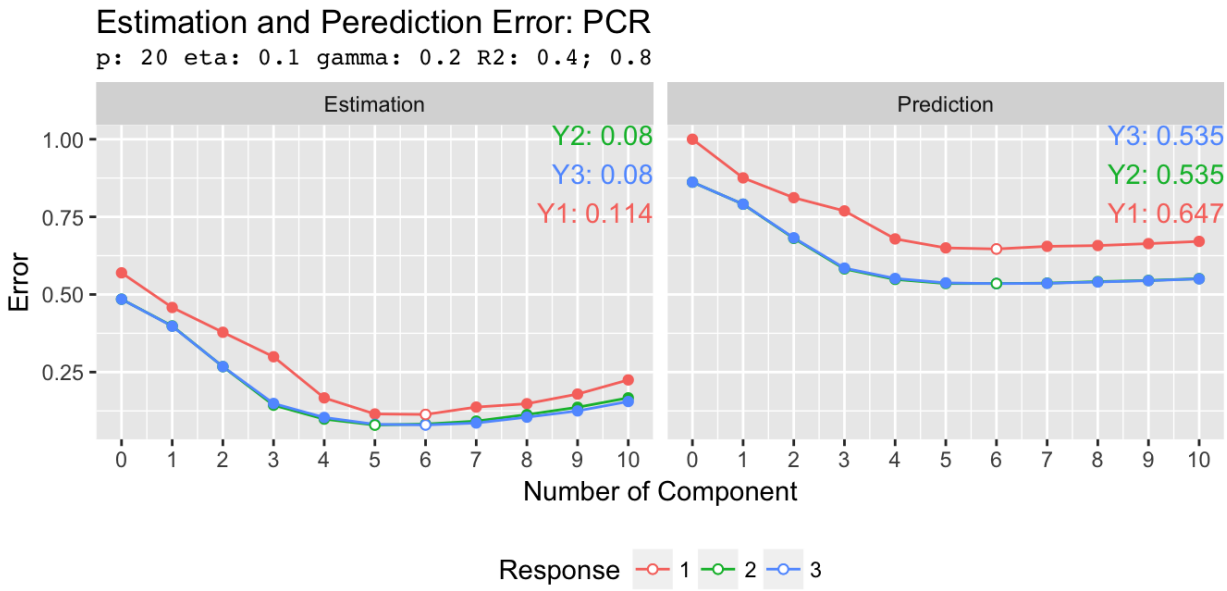
Exploratory Study

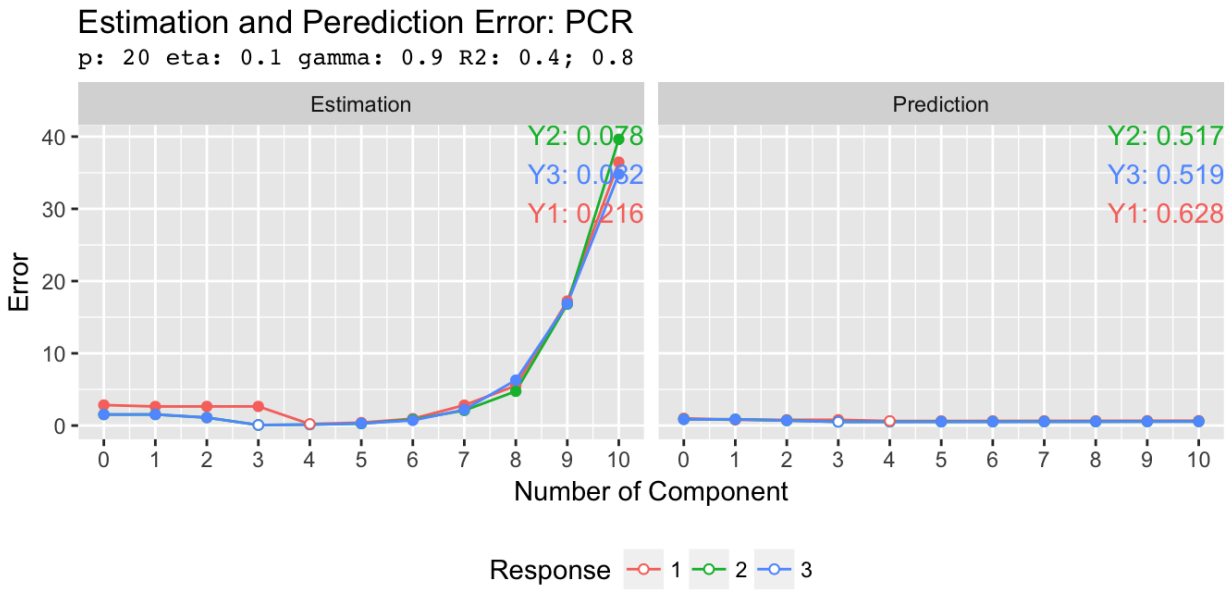
- This section explores the inter-connection between the estimation methods and the properties of data based on regression coefficients
- Our discussion revolve around following factors and their interaction
  - Wide vs Tall predictor matrix



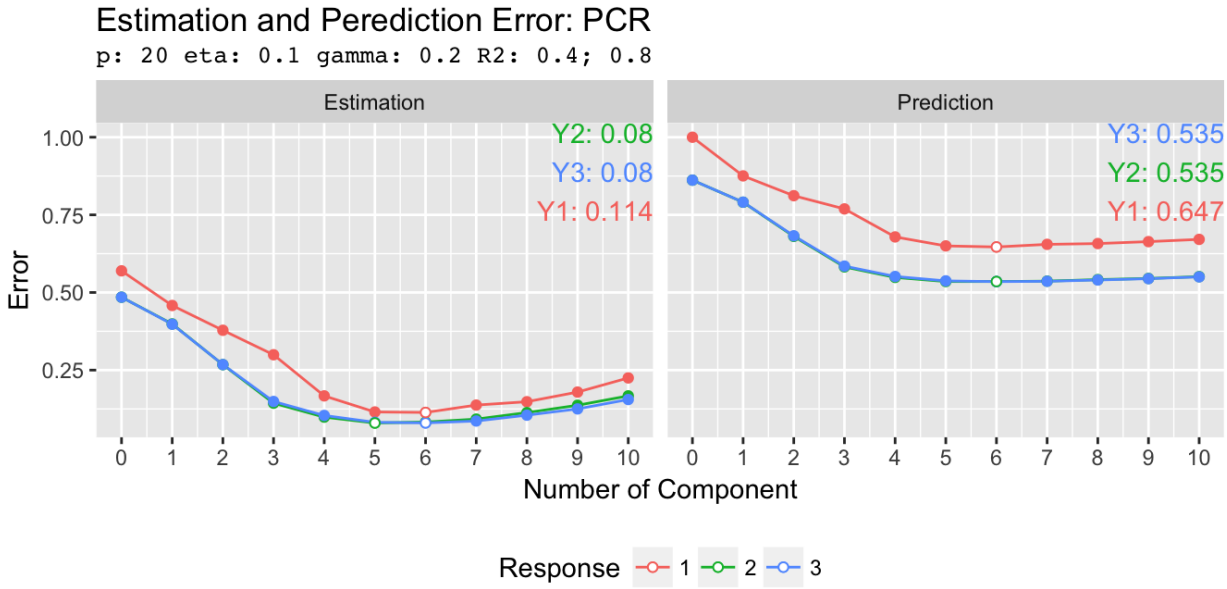


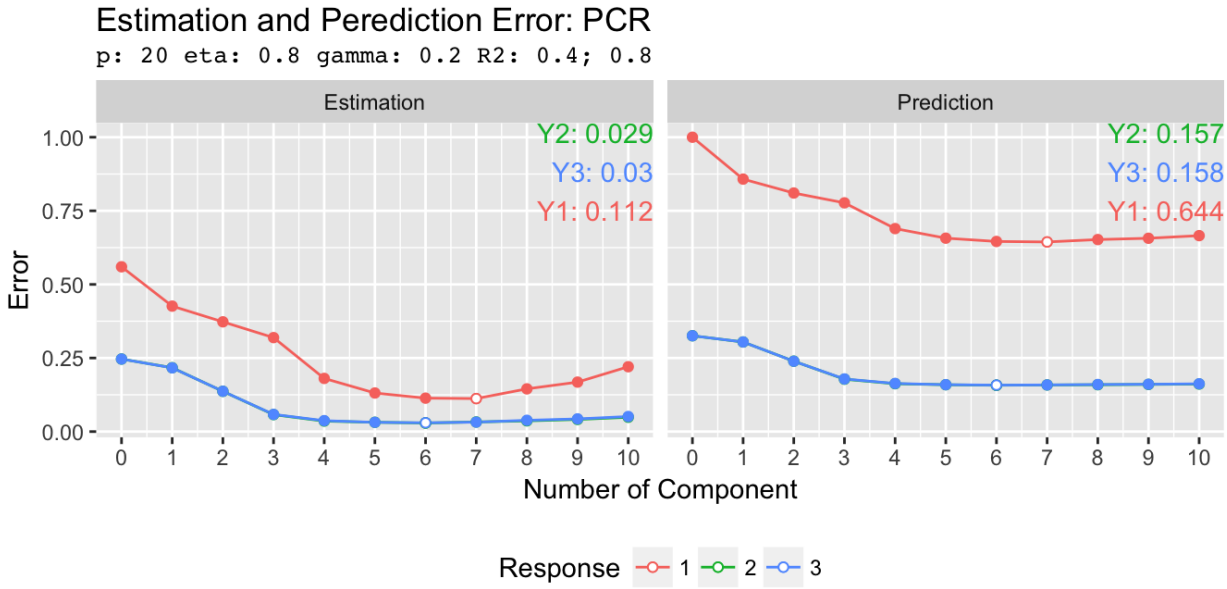
- High vs Low multicollinearity



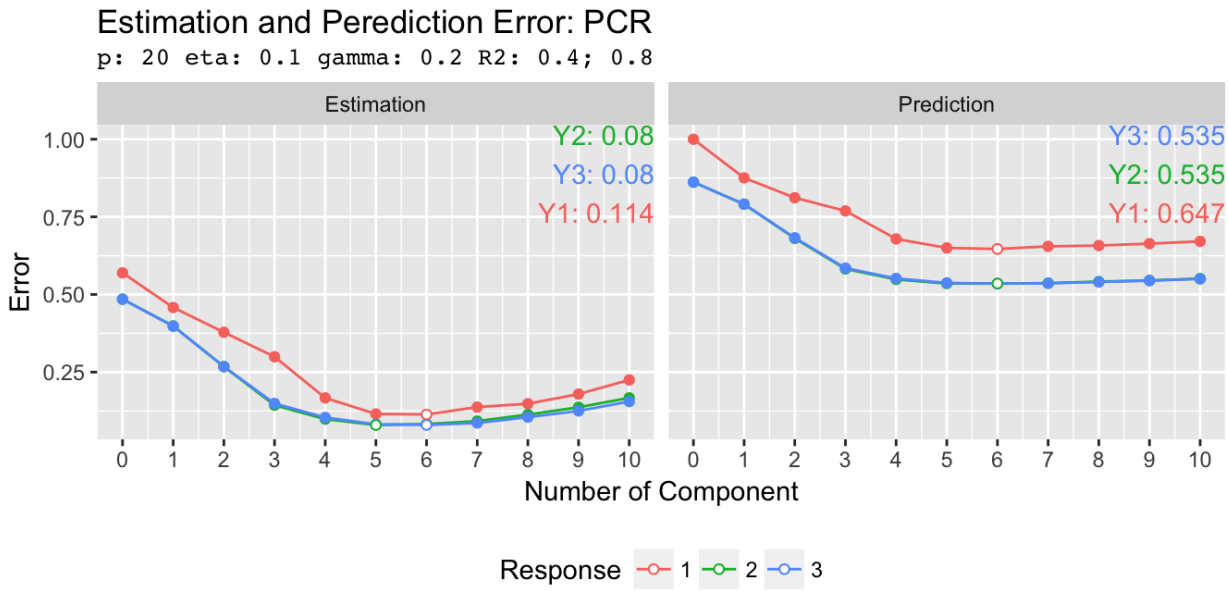


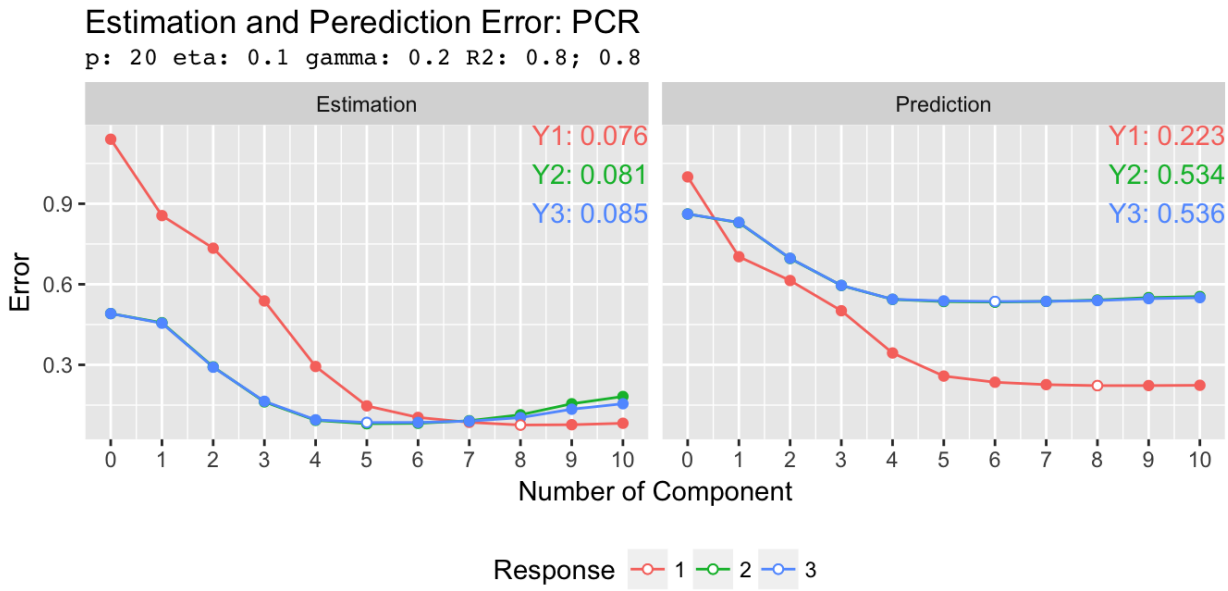
- High vs Low correlation between responses





- Hight vs Low coefficient of determination





Systematic Comparison

- *Should we use MANOVA model or some kind of norm/trace or similar measure for the error and use ANVOA instead?*
- A MANOVA model is used for statistical analysis

$$\texttt{pred\_err}_{ijklm} = \mu + \texttt{p}_i * \texttt{gamma}_j * \texttt{r2}_k * \texttt{method}_l * \texttt{eta}_m + \epsilon_{ijklm}$$

- In the model the prediction error for each of three response variables are used as response variable and following variables (with levels) and their complete interactions are used as predictor variables.
  a. Number of predictor variables (p): 20 and 250
  b. Decay factor of eigenvalues of X (γ): 0.2 and 0.9

c. Decay factor of eigenvalues of Y (η): 0.1 and 0.8
d. Coefficient of Determination (ρ): 0.8, 0.8 and 0.4, 0.8
e. Method of estimation: PCR, PLS1, PLS2, Xenvelope, Yenvelope, Senvelope, Ridge and Lasso
f. Number of tuning Parameters used (as numeric)

- Following is the MANOVA output for estimation error and prediction error models using number of components (tuning parameters) that results minimum error.

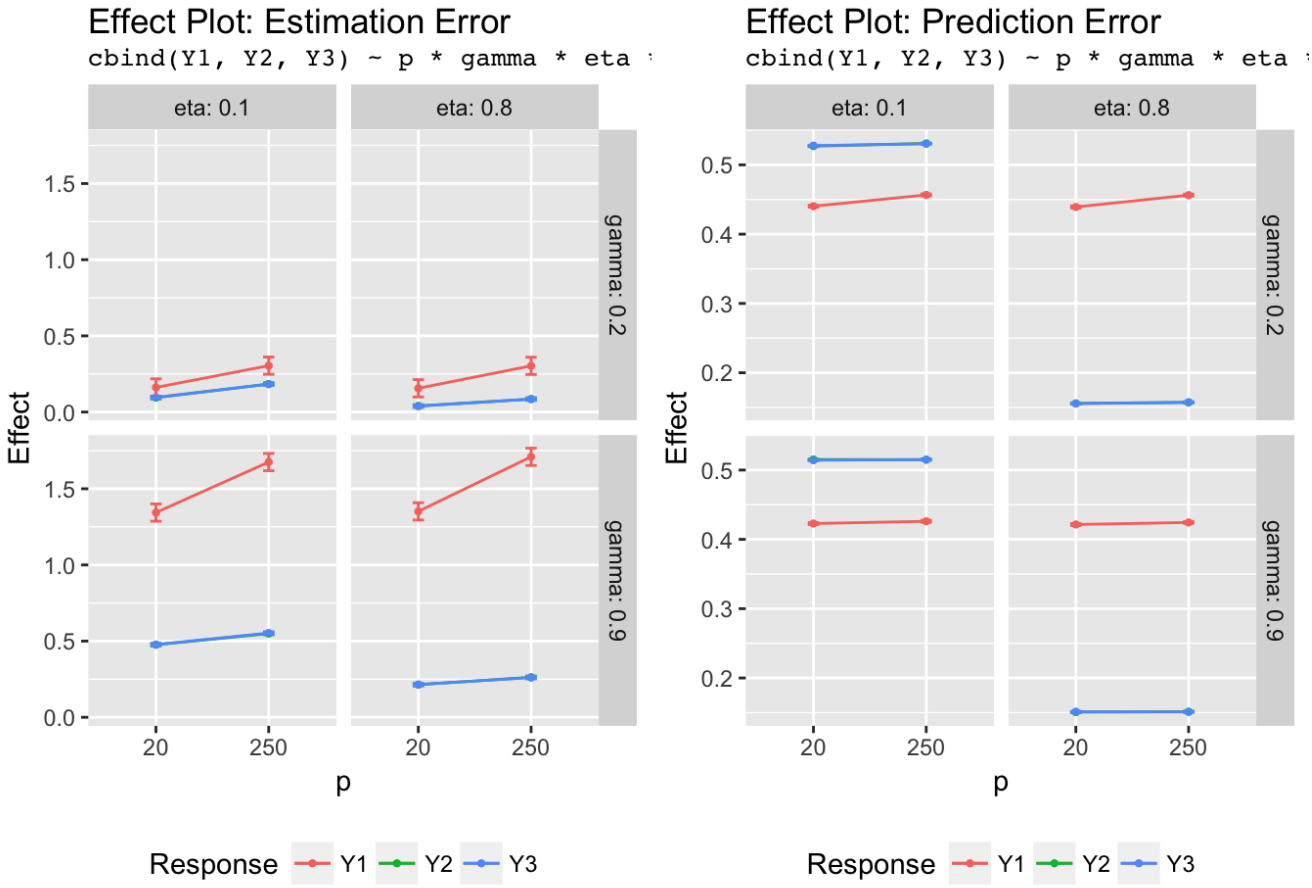
Estimation Error Model:

```
Analysis of Variance Table

                        Df  Pillai approx F num Df den Df    Pr(>F)
(Intercept)              1 0.85222  12052.7      3   6270 < 2.2e-16 ***
p                        1 0.09529    220.1      3   6270 < 2.2e-16 ***
gamma                    1 0.68418   4527.6      3   6270 < 2.2e-16 ***
eta                      1 0.38443   1305.2      3   6270 < 2.2e-16 ***
R2                       1 0.05217    115.0      3   6270 < 2.2e-16 ***
Method                   7 0.88602    375.5     21  18816 < 2.2e-16 ***
p:gamma                  1 0.00396      8.3      3   6270 1.648e-05 ***
p:eta                    1 0.00613     12.9      3   6270 2.187e-08 ***
gamma:eta                1 0.16480    412.4      3   6270 < 2.2e-16 ***
p:R2                     1 0.00577     12.1      3   6270 6.556e-08 ***
gamma:R2                 1 0.04252     92.8      3   6270 < 2.2e-16 ***
eta:R2                   1 0.00009      0.2      3   6270 0.9011113
p:Method                 7 0.21428     68.9     21  18816 < 2.2e-16 ***
gamma:Method             7 0.78221    316.0     21  18816 < 2.2e-16 ***
eta:Method               7 0.40078    138.2     21  18816 < 2.2e-16 ***
R2:Method                7 0.12490     38.9     21  18816 < 2.2e-16 ***
p:gamma:eta              1 0.00060      1.3      3   6270 0.2862542
p:gamma:R2               1 0.00040      0.8      3   6270 0.4739464
p:eta:R2                 1 0.00031      0.7      3   6270 0.5806291
gamma:eta:R2             1 0.00017      0.4      3   6270 0.7874492
p:gamma:Method           7 0.07353     22.5     21  18816 < 2.2e-16 ***
p:eta:Method             7 0.02282      6.9     21  18816 < 2.2e-16 ***
gamma:eta:Method         7 0.23322     75.5     21  18816 < 2.2e-16 ***
p:R2:Method              7 0.01715      5.2     21  18816 1.125e-13 ***
gamma:R2:Method          7 0.09159     28.2     21  18816 < 2.2e-16 ***
eta:R2:Method            7 0.00190      0.6     21  18816 0.9409893
p:gamma:eta:R2           1 0.00003      0.1      3   6270 0.9830458
p:gamma:eta:Method       7 0.00613      1.8     21  18816 0.0112070 *
p:gamma:R2:Method        7 0.00793      2.4     21  18816 0.0003875 ***
p:eta:R2:Method          7 0.00359      1.1     21  18816 0.3696117
gamma:eta:R2:Method      7 0.00307      0.9     21  18816 0.5684497
p:gamma:eta:R2:Method    7 0.00308      0.9     21  18816 0.5651083
Residuals             6272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction Error Model:

```
Analysis of Variance Table

                         Df  Pillai approx F num Df den Df    Pr(>F)
(Intercept)               1 0.99958  5027601      3   6270 < 2.2e-16 ***
p                         1 0.06360      142      3   6270 < 2.2e-16 ***
gamma                     1 0.43493     1609      3   6270 < 2.2e-16 ***
eta                       1 0.99826  1201956      3   6270 < 2.2e-16 ***
R2                        1 0.99118   234856      3   6270 < 2.2e-16 ***
Method                    7 0.87097      367     21  18816 < 2.2e-16 ***
p:gamma                   1 0.03254       70      3   6270 < 2.2e-16 ***
p:eta                     1 0.00083        2      3   6270  0.155656
gamma:eta                 1 0.07579      171      3   6270 < 2.2e-16 ***
p:R2                      1 0.00353        7      3   6270 6.063e-05 ***
gamma:R2                  1 0.04235       92      3   6270 < 2.2e-16 ***
eta:R2                    1 0.00021        0      3   6270  0.726919
p:Method                  7 0.20968       67     21  18816 < 2.2e-16 ***
gamma:Method              7 0.41067      142     21  18816 < 2.2e-16 ***
eta:Method                7 0.14667       46     21  18816 < 2.2e-16 ***
R2:Method                 7 0.26131       85     21  18816 < 2.2e-16 ***
p:gamma:eta               1 0.00200        4      3   6270  0.005629 **
p:gamma:R2                1 0.00669       14      3   6270 3.855e-09 ***
p:eta:R2                  1 0.00027        1      3   6270  0.641880
gamma:eta:R2              1 0.00188        4      3   6270  0.008216 **
p:gamma:Method            7 0.20889       67     21  18816 < 2.2e-16 ***
p:eta:Method              7 0.01058        3     21  18816 1.262e-06 ***
gamma:eta:Method          7 0.03617       11     21  18816 < 2.2e-16 ***
p:R2:Method               7 0.07252       22     21  18816 < 2.2e-16 ***
gamma:R2:Method           7 0.09446       29     21  18816 < 2.2e-16 ***
eta:R2:Method             7 0.00529        2     21  18816  0.043606 *
p:gamma:eta:R2            1 0.00129        3      3   6270  0.044494 *
p:gamma:eta:Method        7 0.01106        3     21  18816 4.137e-07 ***
p:gamma:R2:Method         7 0.04514       14     21  18816 < 2.2e-16 ***
p:eta:R2:Method           7 0.00442        1     21  18816  0.146295
gamma:eta:R2:Method       7 0.00578        2     21  18816  0.020300 *
p:gamma:eta:R2:Method     7 0.00321        1     21  18816  0.509954
Residuals              6272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Study of Effect



Discussion and Conclusion