Simulation6

Mengqi Liu

Aug 10, 2023

Recap

- N: number of samples one time.
- M: number of bins.
- H0: $X \perp\!\!\!\perp Y \mid Z$, H1: $X \not\perp\!\!\!\perp Y \mid Z$
- Methods: ($\tilde{Z}$ is the discretized Z, and the data belonging to the same group share the same $\tilde{Z}$.)
  - "Linear_reg_y": regress Y on $1, X, \tilde{Z}$ and take the *absolute* coefficient of X as the test statistic.
  - "Linear_reg_x": regress X on $1, Y, \tilde{Z}$ and take the *absolute* coefficient of Y as the test statistic.
  - "Double_reg": regress Y on $\tilde{Z}$ and regress X on $1, \tilde{Z}$ separately. Take the *absolute* correlation between residuals from two linear regressions as the test statistic.
  - "Linear_reg_y_z": regress Y on $1, X, Z$ and take the *absolute* coefficient of X as the test statistic.
  - "Linear_reg_x_z": regress X on $1, Y, Z$ and take the *absolute* coefficient of Y as the test statistic.
  - "Double_reg_z": regress Y on Z and regress X on $1, Z$ separately. Take the *absolute* correlation between residuals from two linear regressions as the test statistic.

- $\alpha = 0.05$

- $X = f_x(Z) + \epsilon$, $Y = f_y(Z) + \epsilon$

- Noise ϵ:

  - various a, cor
  - H0:

    - normal: $N(Z, a)$
    - skewed_normal: $N(Z, a)$
  - H1:

    - normal: $N\left([0,0], \begin{pmatrix} a_1 & cor \cdot \sqrt{a_1 a_2} \\ cor \cdot \sqrt{a_1 a_2} & a_2 \end{pmatrix}\right)$
    - skewed_normal: $N\left([0,0], \begin{pmatrix} a_1 & cor \cdot \sqrt{a_1 a_2} \\ cor \cdot \sqrt{a_1 a_2} & a_2 \end{pmatrix}\right)$, skewness = $[5, -5]$
- $N = 100, Z \sim \text{Unif}([0, 10))$, $M \in \{2, 5, 10, 16, 25, 50\}$.

Gains

- When $X|Z$ and $Y|Z$ are both smooth, all methods have valid type-I error and notable power whichever variable is permutated (experiment 1). Moreover, there's no distinct difference generally in power between using Z and $\tilde{Z}$ (experiment 2&experiment 3). It's noteworthy that methods using $\tilde{Z}$ seem to have higher power than methods using Z when cor is relatively small.

- When $X|Z$ and $Y|Z$ are neither smooth, all methods fail in controlling type-I error while "Double_reg" (double regression with Z) is barely acceptable if f_x and f_y are both linear in Z. (experiment 4&experiment 5)

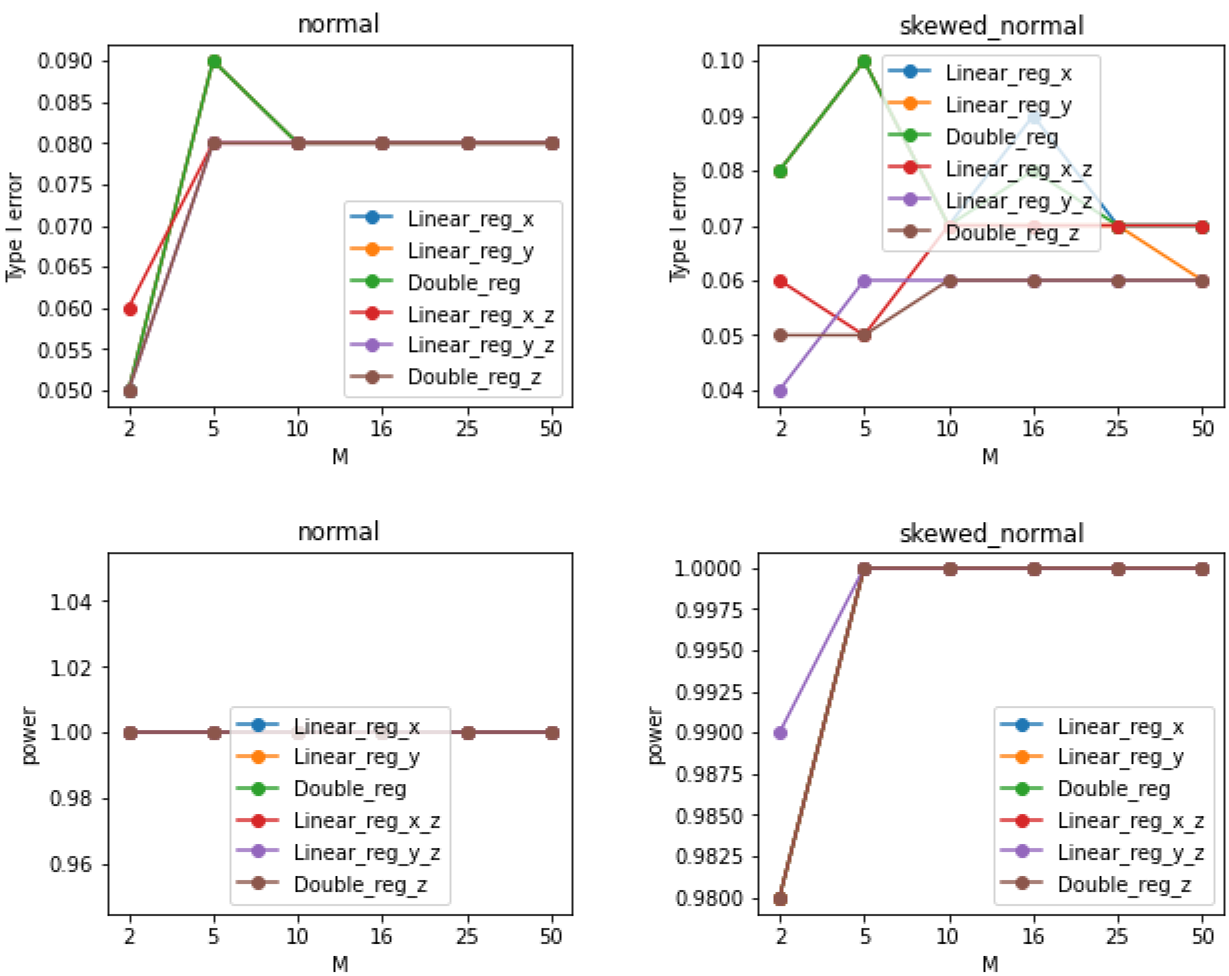
- When $X|Z$ is smooth and $Y|Z$ is not, the choice of permutated variable is very important. All methods perform well if we permute the smooth variable (i.e. X). If we permute the non-smooth one (i.e. Y), "Double_reg" and "Double_reg_z" can both work (experiment 6&experiment 7)

- From all experiments, counter-intuitively, it seems that using $\tilde{Z}$ would be more aggressive (higher type-I and power) than using Z.

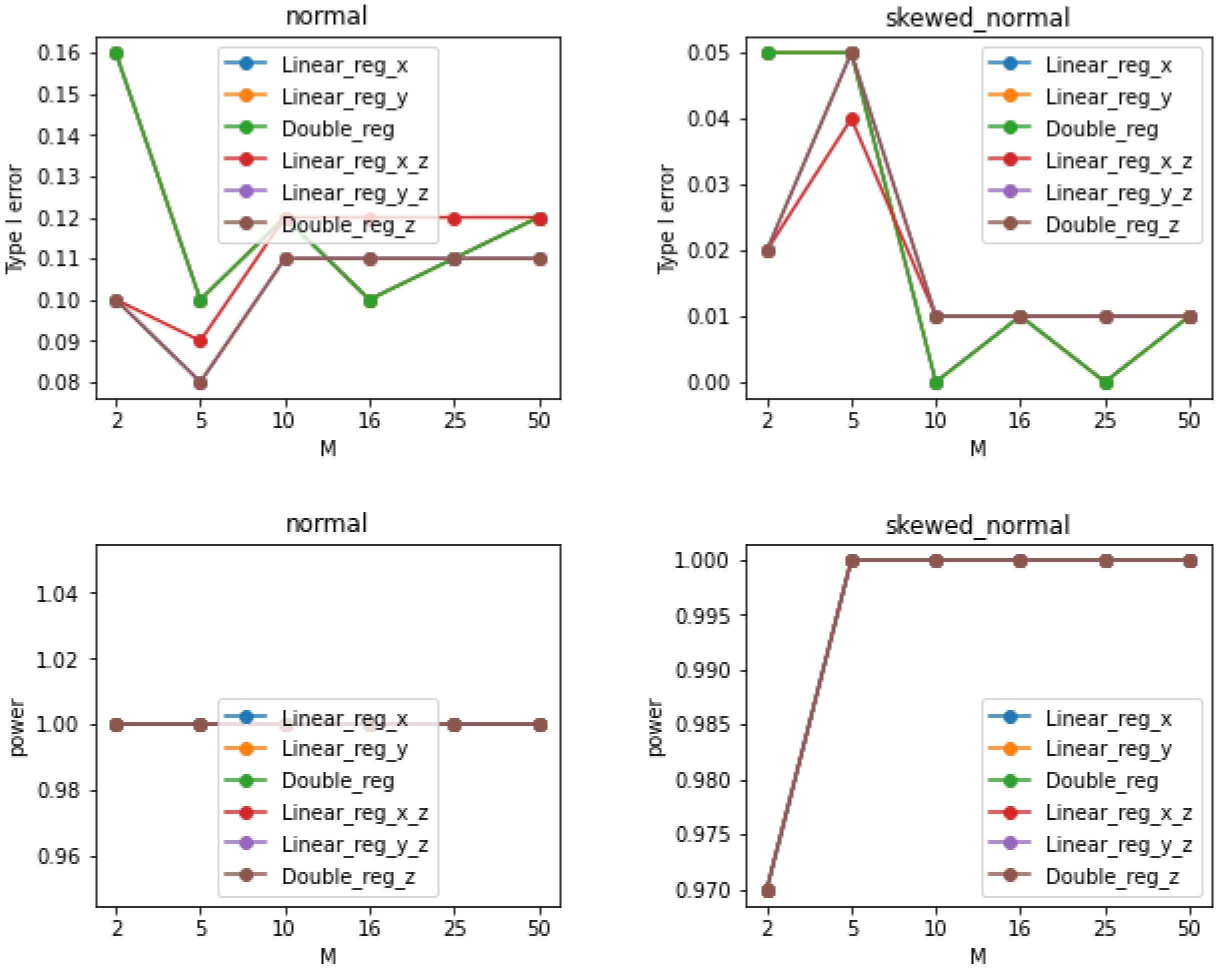
experiment 1

$f_x(Z) = Z$, $f_y(Z) = Z$, $\epsilon_x \sim N(\cdot, 5)$, $\epsilon_y \sim N(\cdot, 5)$, $cor = 0.4$.

- permute Y:

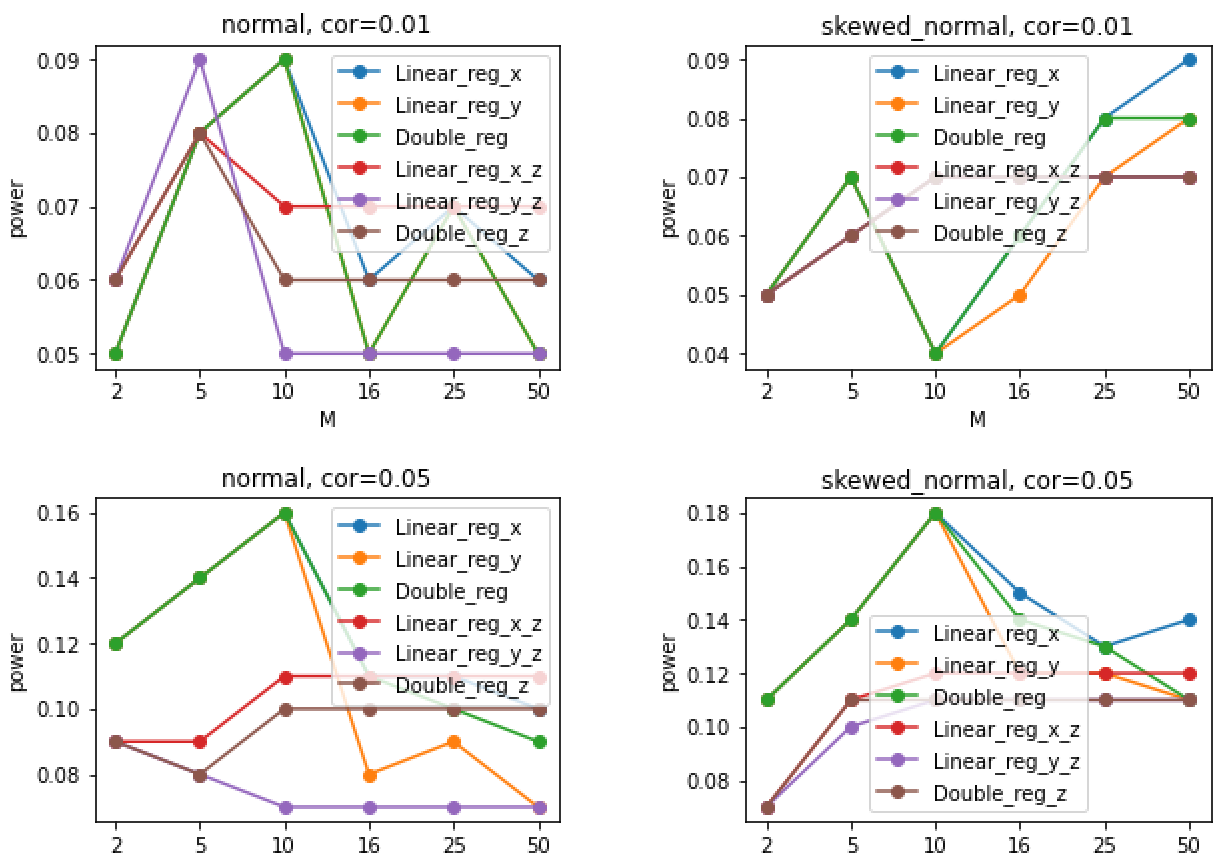


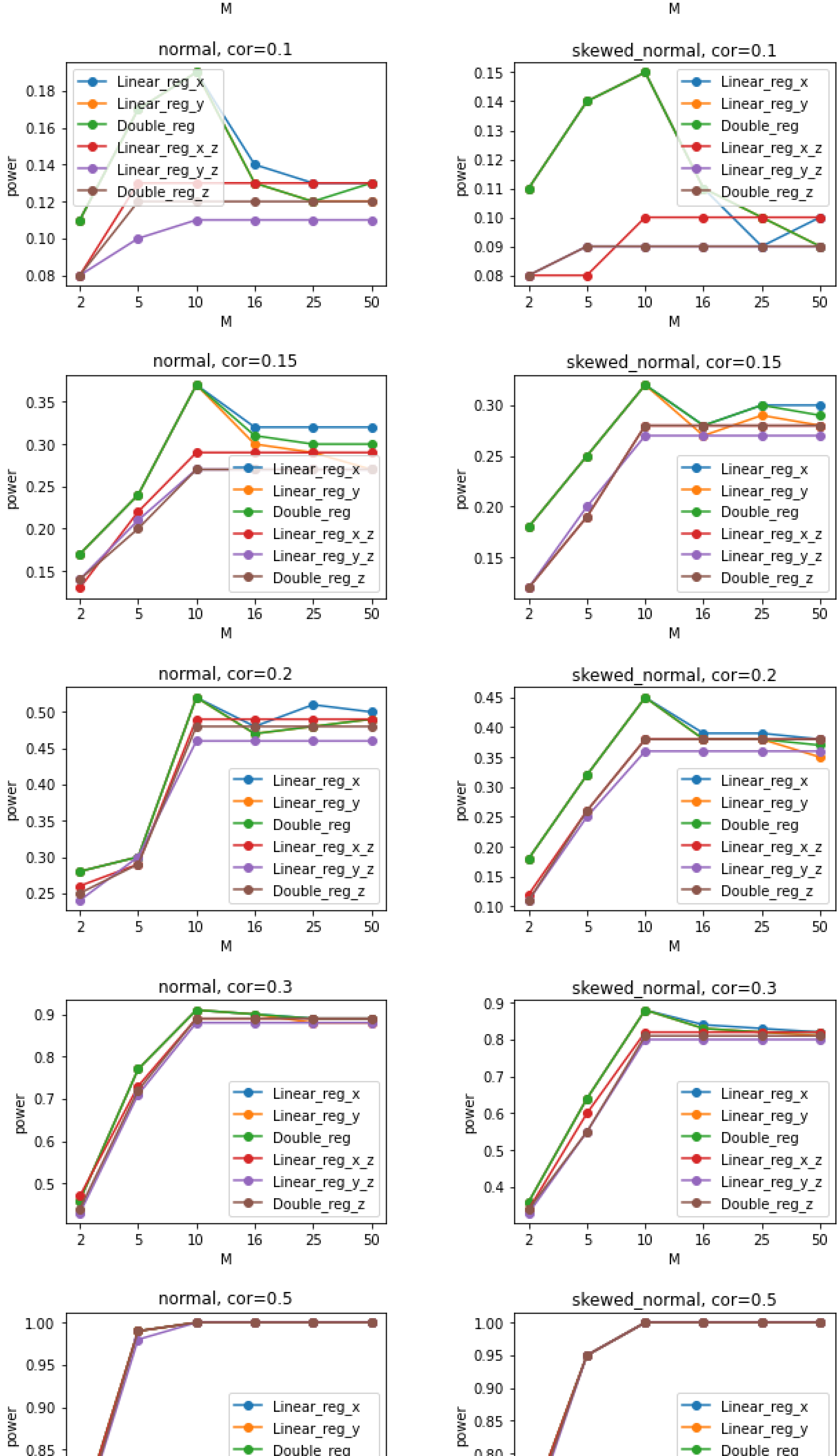
- permute X:

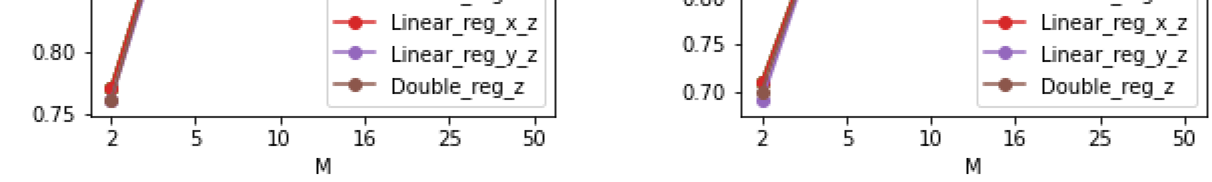


experiment 2

$f_x(Z) = Z$, $f_y(Z) = Z$, $\epsilon_x \sim N(\cdot, 5)$, $\epsilon_y \sim N(\cdot, 5)$, $cor \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5, 0.7\}$ and permute Y.

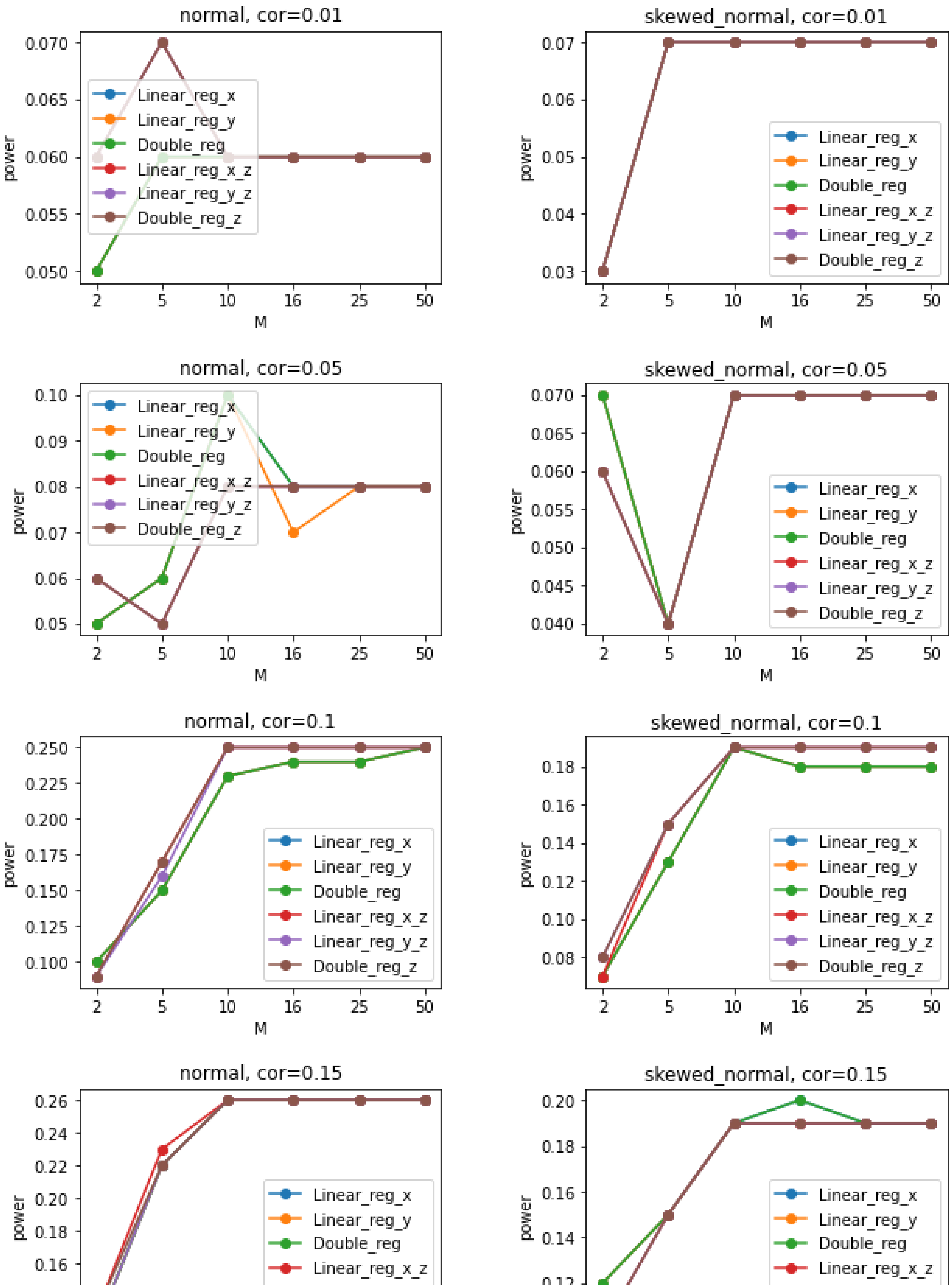


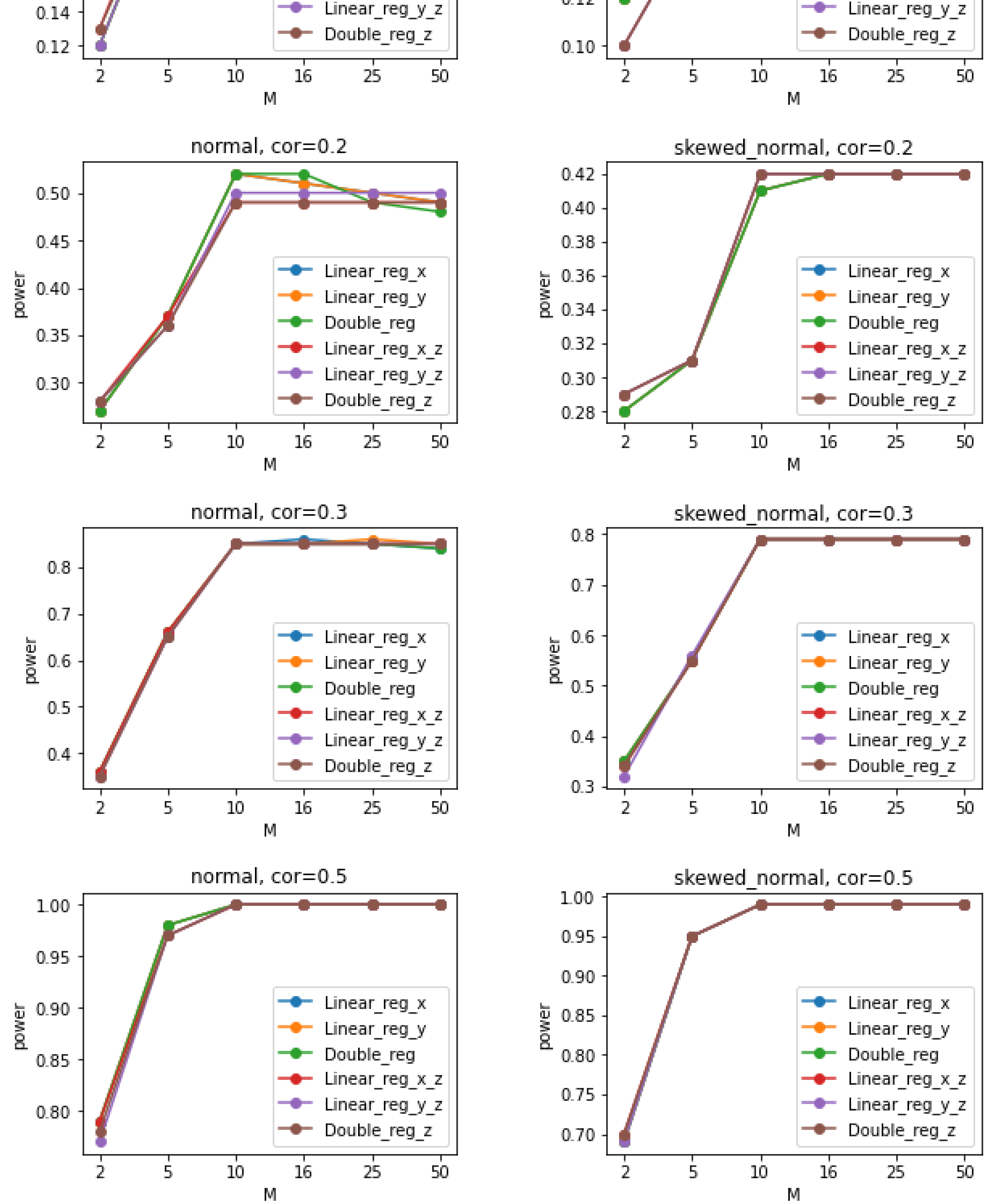

M
M
normal, cor=0.1
skewed_normal, cor=0.1
normal, cor=0.15
skewed_normal, cor=0.15
normal, cor=0.2
skewed_normal, cor=0.2
normal, cor=0.3
skewed_normal, cor=0.3
normal, cor=0.5
skewed_normal, cor=0.5
power
M
2 5 10 16 25 50
Linear_reg_x
Linear_reg_y
Double_reg
Linear_reg_x_z
Linear_reg_y_z
Double_reg_z



experiment 3

$f_x(Z) = \log(Z+1) + 2$, $f_y(Z) = 7 + \sqrt{Z}$, $\epsilon_x \sim N(\cdot, 5)$, $\epsilon_y \sim N(\cdot, 5)$, $cor \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5, 0.7\}$ and permute Y.

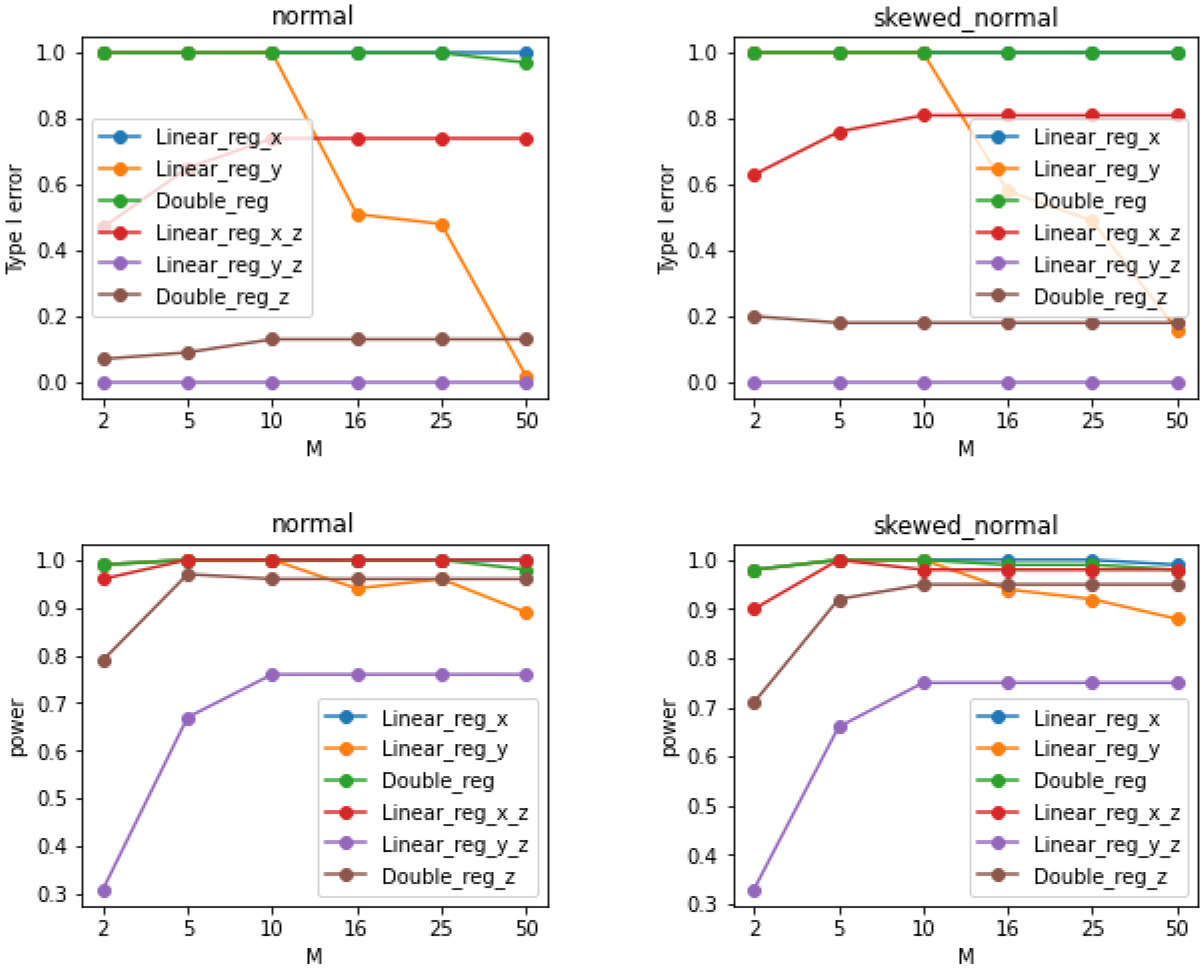




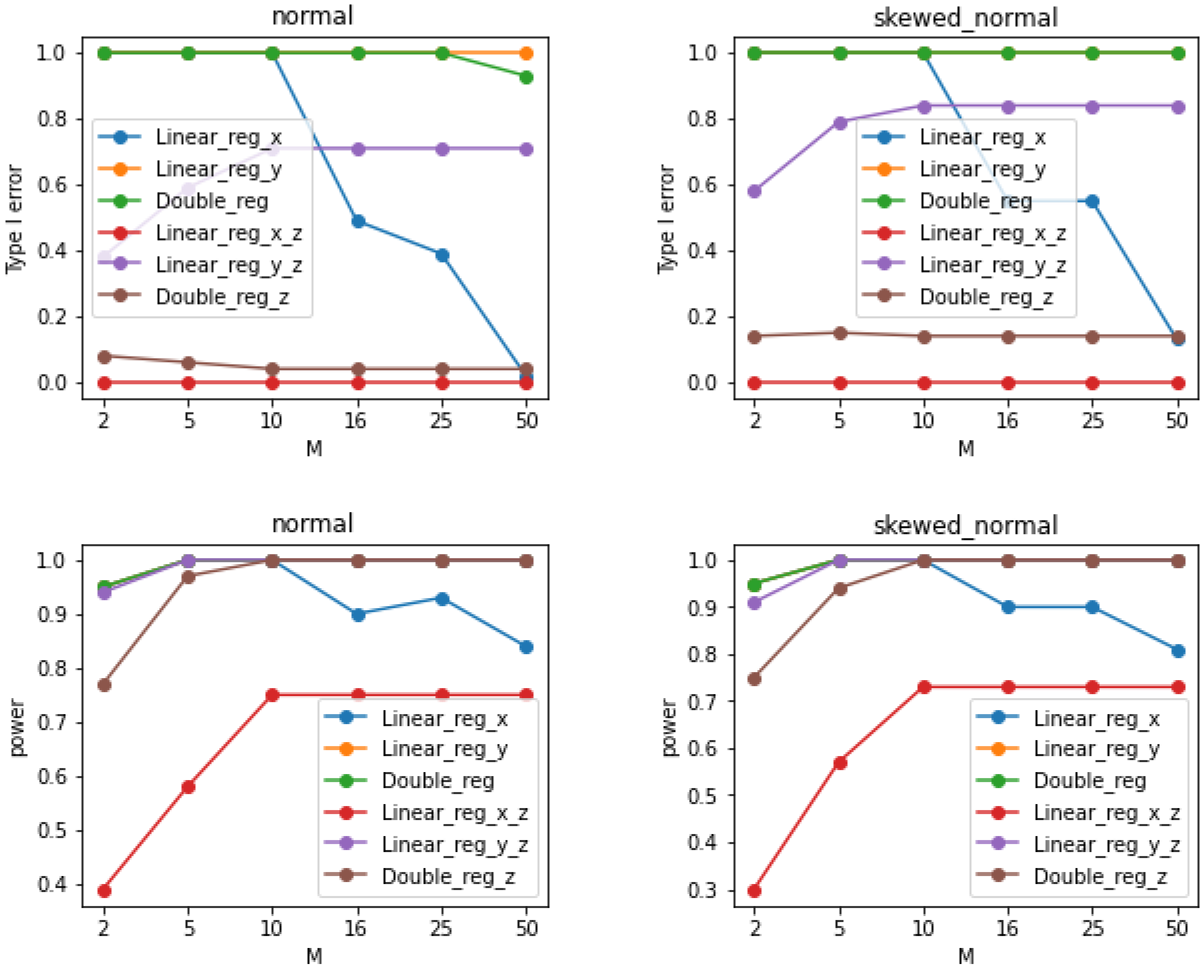
experiment 4

$f_x(Z) = Z$, $f_y(Z) = Z$, $\epsilon_x \sim N(\cdot, 0.1)$, $\epsilon_y \sim N(\cdot, 0.1)$, cor=0.4.

- permute Y:


normal
skewed_normal
Type I error
power
M
2
5
10
16
25
50
Linear_reg_x
Linear_reg_y
Double_reg
Linear_reg_x_z
Linear_reg_y_z
Double_reg_z
0.0
0.2
0.4
0.6
0.8
1.0
0.3
0.4
0.5
0.6
0.7
0.8
0.9

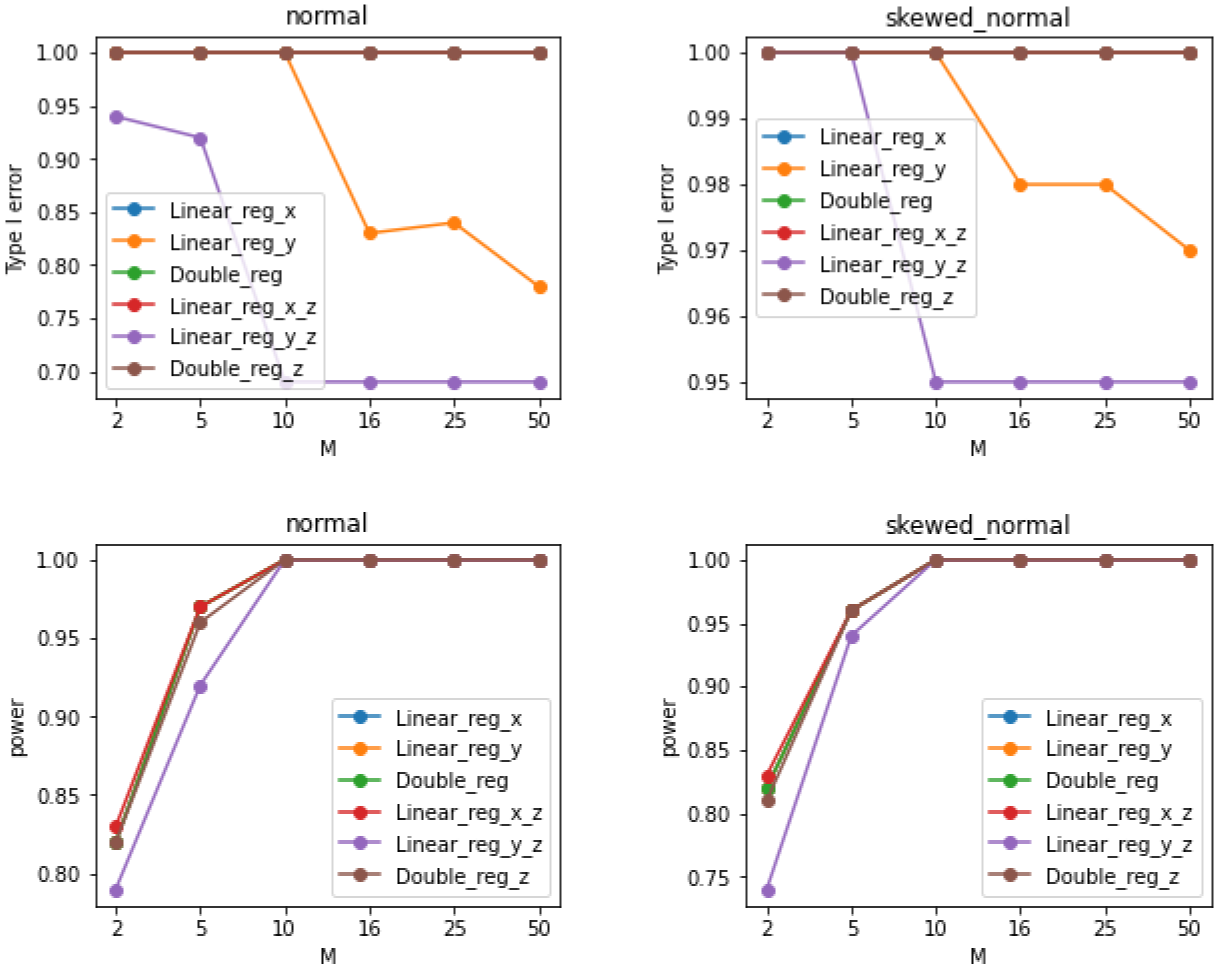
- permute X:



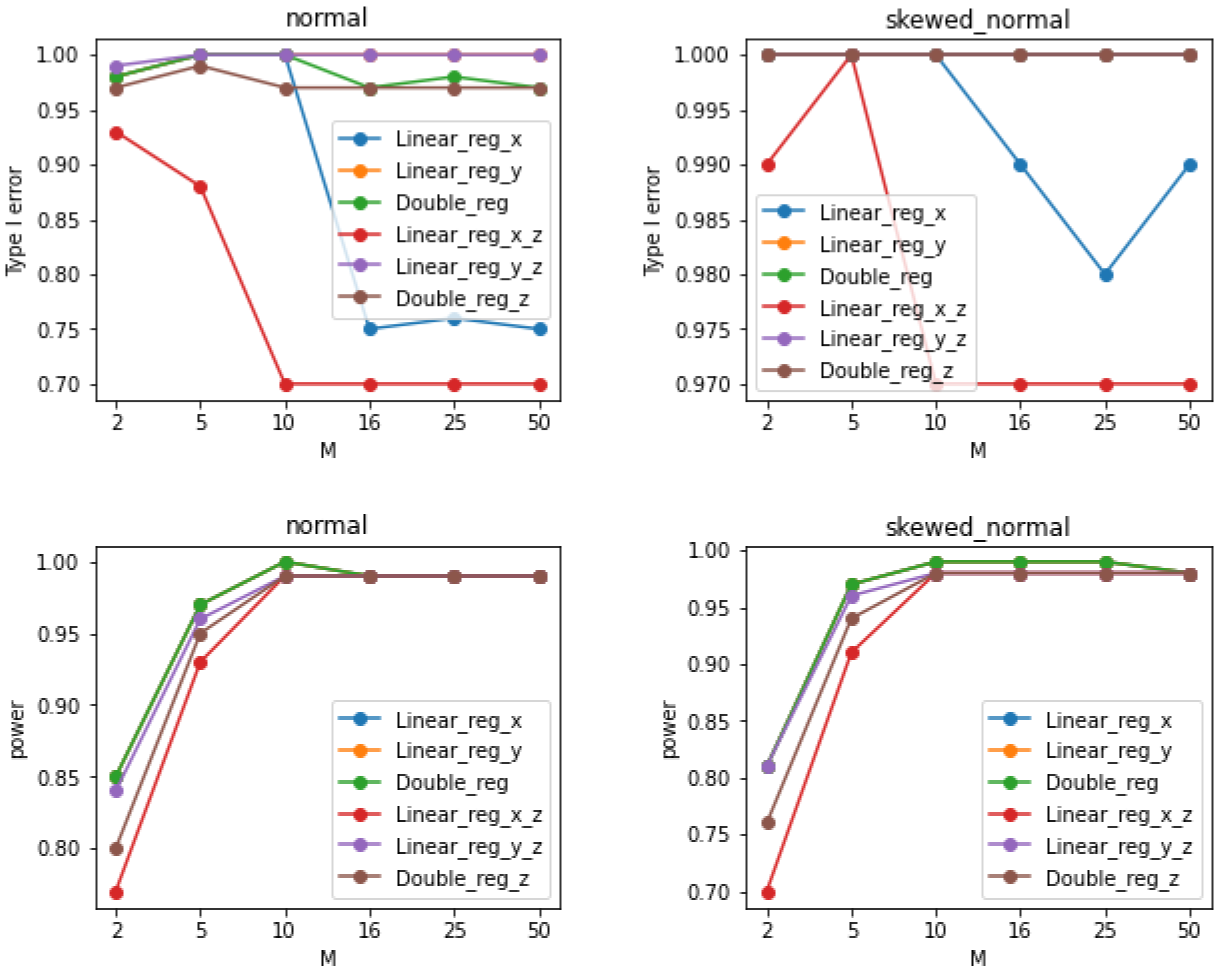
experiment 5

$f_x(Z) = \log(Z+1) + 2$, $f_y(Z) = 7 + \sqrt{Z}$, $\epsilon_x \sim N(\cdot, 0.1)$, $\epsilon_y \sim N(\cdot, 0.1)$, $cor = 0.4$.

- permute Y:


normal
skewed_normal
Type I error
M
power
Linear_reg_x
Linear_reg_y
Double_reg
Linear_reg_x_z
Linear_reg_y_z
Double_reg_z
2
5
10
16
25
50

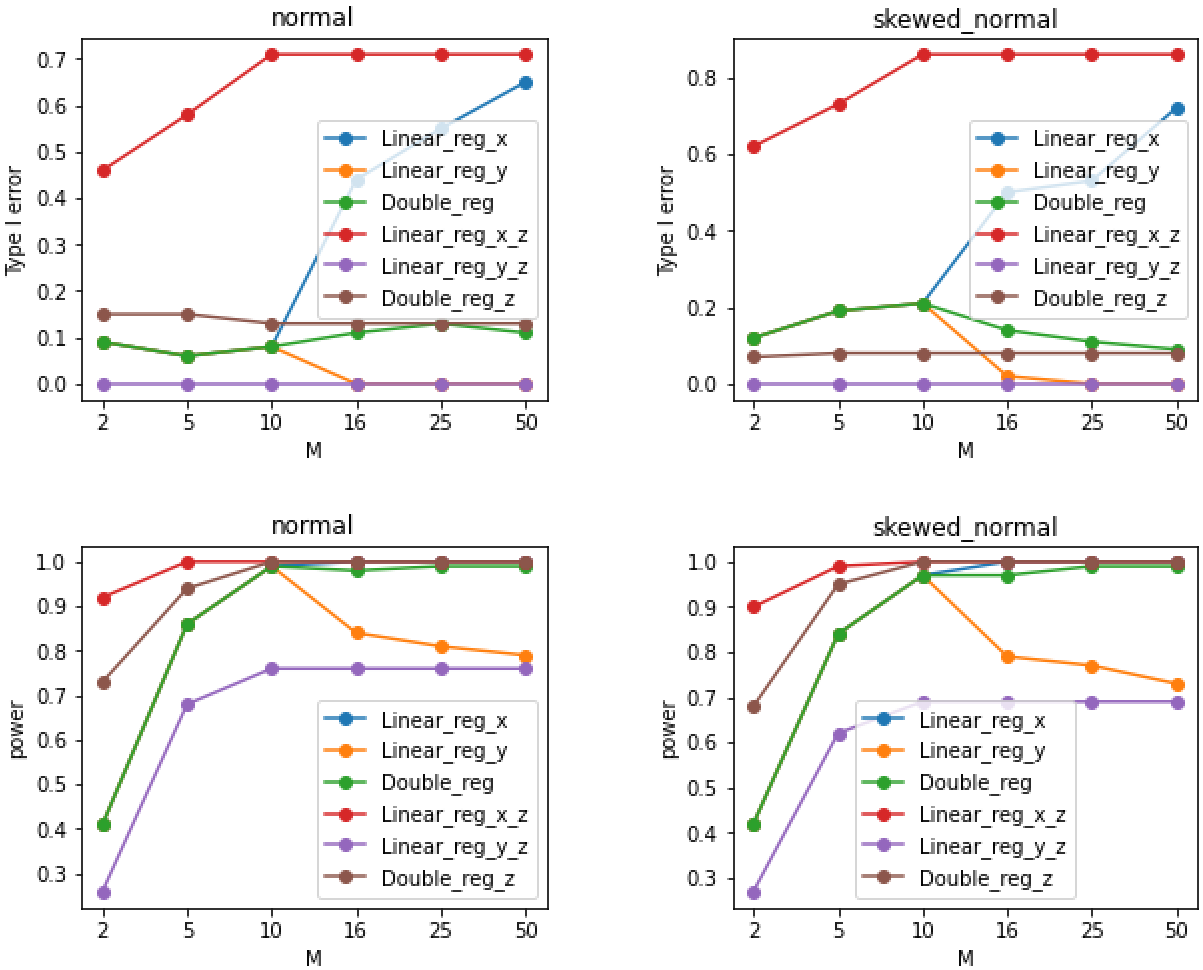
- permute X:



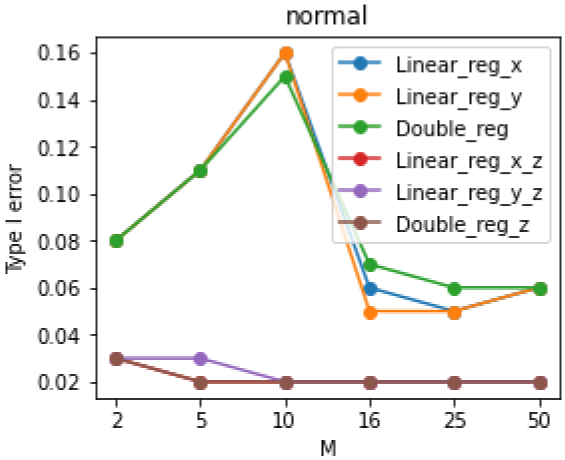
experiment 6

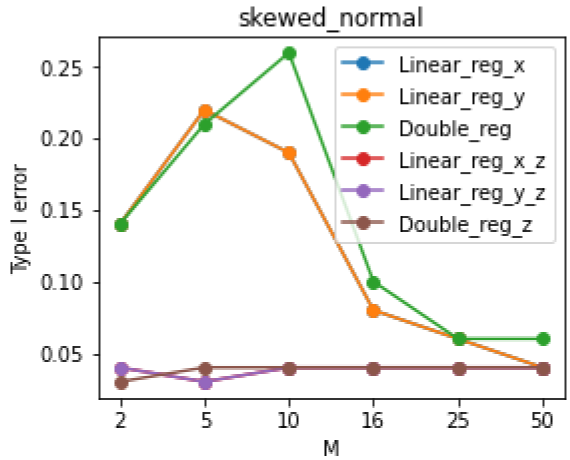
$f_x(Z) = Z$, $f_y(Z) = Z$, $\epsilon_x \sim N(\cdot, 5)$, $\epsilon_y \sim N(\cdot, 0.1)$, $cor = 0.4$.

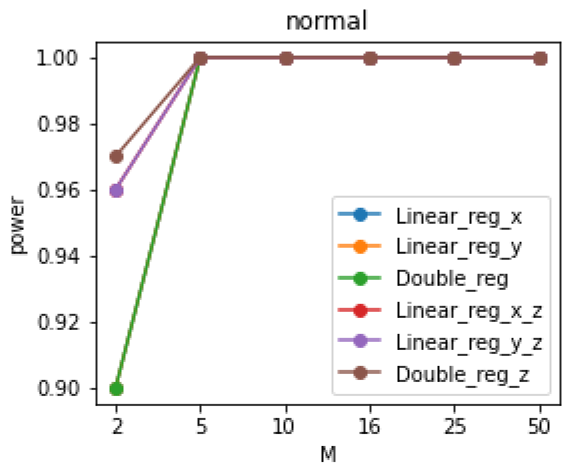
- permute Y:

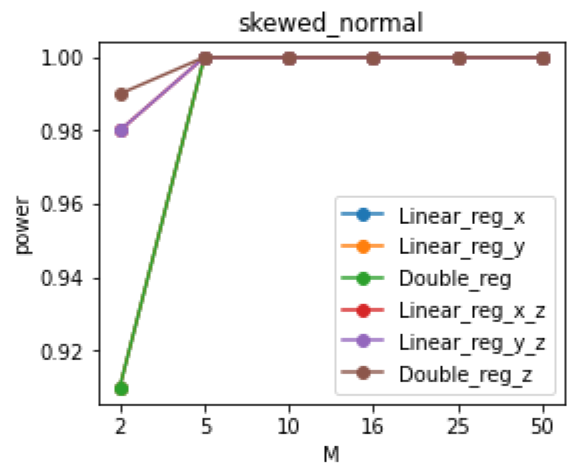

normal
skewed_normal
Type I error
M
Linear_reg_x
Linear_reg_y
Double_reg
Linear_reg_x_z
Linear_reg_y_z
Double_reg_z
normal
skewed_normal
power
M
Linear_reg_x
Linear_reg_y
Double_reg
Linear_reg_x_z
Linear_reg_y_z
Double_reg_z

- permute X:





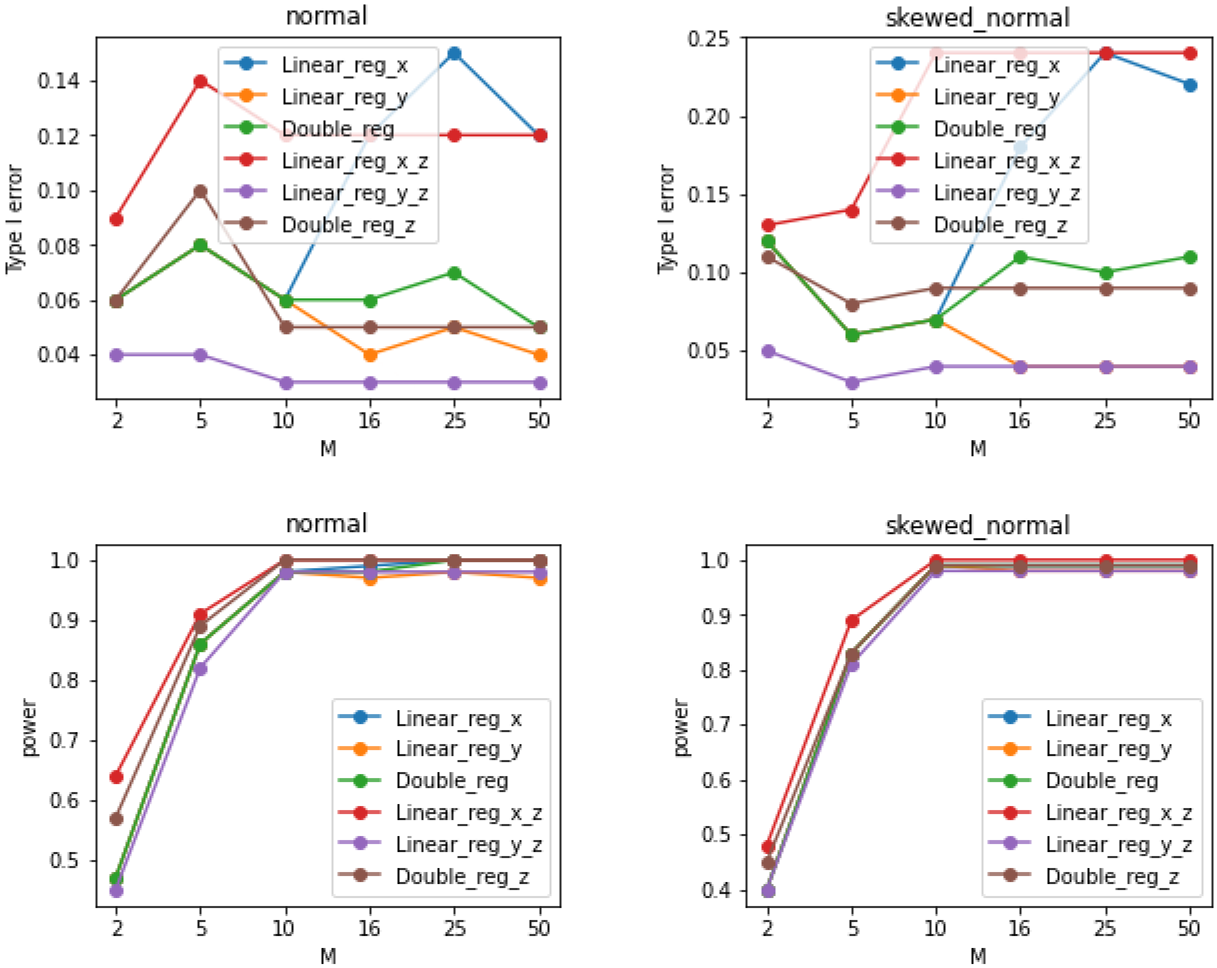




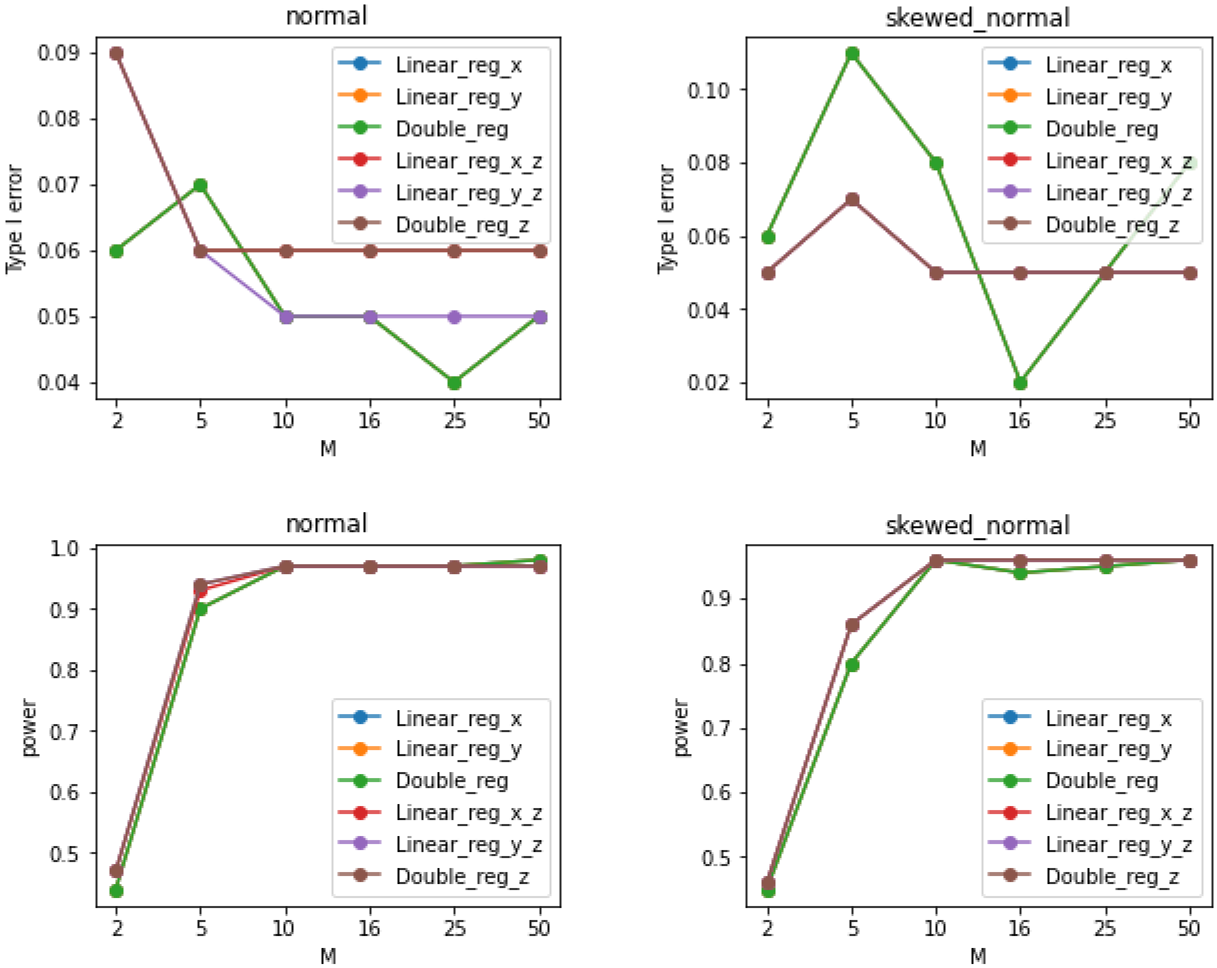
experiment 7

$f_x(Z) = \log(Z+1) + 2$, $f_y(Z) = 7 + \sqrt{Z}$, $\epsilon_x \sim N(\cdot, 5)$, $\epsilon_y \sim N(\cdot, 0.1)$, $cor = 0.4$.

- permute Y:


normal
Type I error
M
Linear_reg_x
Linear_reg_y
Double_reg
Linear_reg_x_z
Linear_reg_y_z
Double_reg_z
skewed_normal
Type I error
M
normal
power
M
skewed_normal
power
M

- permute X:


normal
skewed_normal
Type I error
power
M
Linear_reg_x
Linear_reg_y
Double_reg
Linear_reg_x_z
Linear_reg_y_z
Double_reg_z