

1. [10] I have two identical regressors,  $b = a$ , Consider the LASSO regression loss function,

$$\text{loss}(\hat{\beta}_a, \hat{\beta}_b) = \sum_i (y_i - \hat{y}_i)^2 + \lambda \cdot (|\hat{\beta}_a| + |\hat{\beta}_b|), \quad \hat{y}_i = \hat{\beta}_a a_i + \hat{\beta}_b b_i.$$

- (a) [7] Find the penalized estimates  $\hat{\beta}_a$  and  $\hat{\beta}_b$  for arbitrary  $\lambda > 0$ .

In the regression  $\hat{y}_i = \hat{\gamma}_1 a_i$  the OLS estimator  $\hat{\gamma}_1$  is equal to 2.

- (b) [3] What are the approximate values of  $\hat{\beta}_a$  and  $\hat{\beta}_b$  in the penalized model for small  $\lambda \approx 0$ ?
2. [10] Elon Musk studies the model  $y_i = \beta_0 + \beta_1 a_i + \beta_2 b_i + u_i$ . He has 500 observations and he knows the heteroskedasticity form up to an unknown constant. To obtain the most efficient unbiased estimators  $\hat{\beta}$  he has multiplied each observation by  $(a_i/b_i)$ .

- (a) [5] What is the variance  $\text{Var}(u_i | X)$  supposed by Elon Musk?

- (b) [5] To check whether the heteroskedasticity is really present Elon Musk estimated the regression

$$\hat{v}_i = \hat{\gamma}_0 + \hat{\gamma}_1 a_i + \hat{\gamma}_2 b_i + \hat{\gamma}_3 a_i b_i, \quad R^2 = 0.05,$$

where  $v_i$  are the squared residuals  $\hat{u}_i^2$  from original model.

What is the statistical conclusion about the heteroskedasticity presence?

Critical values for 5% significance level:  $\chi_1^2 = 3.84$ ,  $\chi_2^2 = 5.99$ ,  $\chi_3^2 = 7.81$ ,  $\chi_{498}^2 = 549$ ,  $\chi_{499}^2 = 551$ ,  $\chi_{500}^2 = 553$ .

3. [10] Donald Trump has 403 observations. He estimated the first simple regression:

$$\hat{x}_i = 2 - 2w_i, \quad R^2 = 0.95, \quad \text{SST} = 100 \quad (\text{Regression A}).$$

Then he estimated the second regression,

$$\hat{w}_i = \hat{\gamma}_0 + \hat{\gamma}_1 x_i, \quad \text{SST} = 200 \quad (\text{Regression B}).$$

- (a) [2] Find  $R^2$  in the regression B.

- (b) [2] Find  $\hat{\gamma}_1$  in the regression B.

Then he estimated the third regression,

$$\hat{y}_i = 2 + 3x_i + 5w_i, \quad R^2 = 0.6, \quad \text{SST} = 500 \quad (\text{Regression C}).$$

- (c) [3] Find the variance inflation factors of  $x_i$  and  $w_i$  in the regression C.

- (d) [3] Find the 95% confidence interval for  $\partial y / \partial x$ .

4. [10] All regressors in  $X$  matrix are standardized. There are 3 regressors and 10000 observations. The sample variances of the principal components are 10, 1, 0.2.

The second regressor can be calculated from principal components as  $0.2pc_1 + 0.3pc_2 + \alpha pc_3$ .

- (a) [3] Reconstruct the  $D$  matrix in the SVD  $X = UDV^T$ .
- (b) [3] How much variance in % does the second principal component explain?
- (c) [1] What are the possible values of  $\alpha$ ?
- (d) [3] Reconstruct the  $V$  matrix as much as possible.
5. [10] Consider the model  $y_i = \beta_0 + \beta_1 x_i + u_i$  with independent and identically distributed observations. The regressor  $x_i$  is correlated with  $u_i$  and with variable  $z_i$ . The variable  $z_i$  is uncorrelated with  $u_i$ .

The variable  $z_i$  is binary and takes only values 0 or 1.

The regressor  $x_i$  is binary and all observations may be divided into four groups:

	$x = 0$	$x = 1$
$z = 0$	$\bar{y}_{00} = 20, n_{00} = 300$	$\bar{y}_{01} = 20, n_{01} = 500$
$z = 1$	$\bar{y}_{10} = 15, n_{10} = 400$	$\bar{y}_{11} = 35, n_{11} = 200$

Here  $\bar{y}_{zx}$  and  $n_{zx}$  are average values of  $y_i$  and the number of observations for each group.

- (a) [8] Find the instrumental variable estimate  $\hat{\beta}_1$ .
- (b) [2] Will the estimate change if the labels  $z = 0$  and  $z = 1$  will be switched?
6. (based on past LSE exams) Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \dots, n.$$

We assume that the errors  $\{u_i\}_{i=1}^n$  are independent random variables with zero mean.

The regressor  $\{x_i\}_{i=1}^n$  is stochastic and Gauss-Markov conditions are satisfied. Under these conditions, the OLS estimator for  $\beta_1$ ,  $\hat{\beta}_1$ , is conditionally unbiased. You are not asked to derive  $\hat{\beta}_1$ .

- (a) [2] Explain the concept of unbiasedness of an estimator.
- (b) [5] Let us consider two other estimators for the slope  $\beta_1$ :

$$\hat{\beta}_1^\circ = \frac{\sum_{i=1}^n (z_i - \bar{z}) y_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \quad \text{and} \quad \hat{\beta}_1^* = \frac{\sum_{i=1}^n (z_i - \bar{z}) y_i}{\sum_{i=1}^n (z_i - \bar{z}) z_i}$$

where  $z_i = x_i^3$  for all  $i$  and  $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$ . Please indicate whether  $\hat{\beta}_1^\circ$  and  $\hat{\beta}_1^*$  are conditionally unbiased estimators for  $\beta_1$ . Clearly show your derivations.

- (c) [3] Briefly indicate how you would choose between the three estimators,  $\hat{\beta}_1$ ,  $\hat{\beta}_1^\circ$ , and  $\hat{\beta}_1^*$ .