Math 583 HW 4 2023 Due Wednesday, Sept. 20.

A) Suppose $\boldsymbol{\Sigma_x} = diag(1, 2, 3, 4)$, $Y_i = Y_i|\boldsymbol{x}_i = x_{i3} + x_{i4} + e_i$ where the e_i are iid and $e_i \perp\!\!\!\perp \boldsymbol{x}_i$. Consider the population OLS and OPLS MLR models. Then $Y = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ where $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Assume

$$\begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} \sim N_{p+1}\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu_x} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\Sigma_x} \end{pmatrix} \right).$$

Then $Y|\boldsymbol{x} \sim Y|\boldsymbol{\beta}_{OLS}^T\boldsymbol{x}$.
a) What is $\boldsymbol{\beta}$?
b) Find $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.
c) Find $\lambda = \lambda_{OPLS}$.
d) Find $\boldsymbol{\beta}_{OPLS}$.
e) Find $V(Y|\boldsymbol{x}) = V(Y|\boldsymbol{\beta}_{OLS}^T\boldsymbol{x})$.
f) Find $V(Y|\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{x})$.

Hints: see the example done in class on notes 24)-25). For e) and f) $\sigma_Y^2 = \boldsymbol{\Sigma}_Y$ is unknown, so $V(Y|\boldsymbol{x}) = \sigma_Y^2 - a$ and $V(Y|\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{x}) = \sigma_Y^2 - d$. Find a and d.

B) Suppose $\hat{\boldsymbol{\beta}}_E$ estimates $\boldsymbol{\beta}_E = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ and the population OLS estimator $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}(\boldsymbol{u}, Y)$ exists (iid cases). Then $\boldsymbol{\beta}_E = \boldsymbol{C}\boldsymbol{\beta}_{OLS}$. Find $\boldsymbol{C}$.

C) We will study the relationship between Y = the *number of women married to civilians* in the district with the predictors x_1 = constant, $x_2 = pop$ = the *population of the district in 1843*, $x_3 = mmen$ = the *number of married civilian men* in the district, x_4 = *mmilmen* = *number of married men in the military* in the district, and $x_5 = milwmn$ = the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. The output below is for OLS forward selection. The confidence intervals were made using the bootstrap (not yet covered). The asterisk * means the variable is in the model.

```
              Estimate  Std.Err   95% shorth CI
Intercept     241.5445  190.7426  [ -218.40, 652.152]
pop=x2          0                 [ -0.0053,  0.0059]
mmen=x3         1.0010    0.0002  [  0.9687,  1.0310]
mmilmen=x4      0                 [ -0.5951,  7.9475]
milwmn=x5       0                 [ -8.7004,  0.5996]
Selection Algorithm: forward
          pop mmen mmilmen milwmn
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
```

a) What is the value of $C_p(I_{min})$ and what is $\hat{\boldsymbol{\beta}}_{I_{min},0}$? (Include the constant corresponding to the intercept x_1.)

b) Which variables, if any, have confidence intervals that do not contain 0? (Hence the variables are needed in the MLR model given that the other variables are in the model.)

c) List the variables, including a constant, that model 3 contains.

Note: Some output uses TRUE instead of an asterisk * (true that the model contains the variable) and FALSE instead of a blank " ". Other variants also occur.

D) Consider the data set in C) with $Y =$ the *number of women married to civilians* in the district, but now let $x_1 = pop =$ the *population of the district in 1843*, $x_2 = mmen =$ the *number of married civilian men* in the district, $x_3 = mmilmen =$ *number of married men in the military* in the district, and $x_4 = milwmn =$ the *number of women married to husbands in the military* in the district. The output for the marginal simple linear regression of Y on x_i is shown below for $i = 1, 2, 3, 4$.

a)What is $\hat{\boldsymbol{\beta}}_{MMLE}$?

Hint: for x_1, the output gives $\hat{\alpha}_1 = -7908.83$ and $\hat{\beta}_1 = 0.178206$. Similarly, find $\hat{\beta}_i$ for $i = 2, 3, 4$ and stack into a vector.

b) Does it make sense that $\hat{\boldsymbol{\beta}}_2 \approx 1$?

```
Label      Estimate        Std. Error    t-value    p-value
Constant  -7908.83         2622.01        -3.016     0.0060
x1         0.178206        0.00412413     43.211     0.0000

Label      Estimate        Std. Error    t-value    p-value
Constant   241.544         190.743         1.266     0.2175
x2         1.00097         0.00180451    554.702     0.0000

Label      Estimate        Std. Error    t-value    p-value
Constant   60931.0         13866.0         4.394     0.0002
x3         48.5630         15.6793         3.097     0.0049

Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   60745.3         13606.7         4.464     0.0002
x4         51.5514         16.1923         3.184     0.0040
```