`Math 583    HW 8 Fall 2020        Due Friday, Oct. 30.`

Quiz 8 on Wednesday will have problems like this homework. Final: Monday, Dec. 7, 8-10 AM. Problem numbers are from Olive (2020). Do the source commands from homework 4.

A) Let the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has full rank p, $E(\boldsymbol{e}) = \boldsymbol{0}$ and $Cov(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$. Then for a large class of iid error distributions, what is the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$? Hint: use the least squares central limit theorem.

7.1. For the output below, an asterisk means the variable is in the model. All models have a constant, so model 1 contains a constant and mmen.

a) List the variables, including a constant, that models 2, 3, and 4 contain.

b) The term out$cp lists the C_p criterion. Which model (1, 2, 3, or 4) is the minimum C_p model I_{min}?

c) Suppose $\hat{\boldsymbol{\beta}}_{I_{min}} = (241.5445, 1.001)^T$. What is $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$?

```
Selection Algorithm: forward #output for Problem 7.1
         pop mmen mmilmen milwmn
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000

      large sample full model inference
      Est.    SE  t   Pr(>|t|)   nparboot      resboot
int -1.249 0.838 -1.49 0.14 [-2.93,-0.093][-3.045,0.473]
L   -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
logW 0.130 0.374  0.35 0.73 [-0.457,0.829][-0.703,0.890]
H    0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.016]
logS 0.640 0.169  3.80 0.00 [ 0.244,1.040][ 0.336,1.012]
```

7.2 Consider the above output for the OLS full model. The column *resboot* gives the large sample 95% CI for β_i using the shorth applied to the $\hat{\beta}^*_{ij}$ for $j = 1, ..., B$ using the residual bootstrap. The standard large sample 95% CI for β_i is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Hence for β_2 corresponding to L, the standard large sample 95% CI is $-0.001 \pm 1.96(0.002) = -0.001 \pm 0.00392 = [-0.00492, 0.00292]$ while the shorth 95% CI is $[-0.005, 0.004]$.

a) Compute the standard 95% CIs for β_i corresponding to log(W), H, and log(S). Also write down the shorth 95% CI. Are the standard and shorth 95% CIs fairly close?

b) Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If the corresponding 95% CI for β_i does not contain 0, then reject H_0 and conclude that the predictor variable X_i is needed in the MLR model. If 0 is in the CI then fail to reject H_0 and conclude that the predictor variable X_i is not needed in the MLR model given that the other predictors are in the MLR model. Which variables, if any, are needed in the MLR model? Use the standard CI if the shorth CI gives a different result. The nontrivial predictor variables are L, log(W), H, and log(S).

7.4. Suppose the full model has p predictors including a constant. Let submodel I have k predictors. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is for the full model. Since $F_I \geq 0$, $C_p(I_{min}) \geq -p$ and $C_p(I) \geq -p$. Assume the full model is one of the submodels considered. Then $-p \leq C_p(I_{min}) \leq p$. Let $\boldsymbol{r}$ be the residual vector for the full model and $\boldsymbol{r}_I$ that for the submodel. Then the correlation

$$corr(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}}.$$

a) Show $corr(r, r_{I_{min}}) \to 1$ as $n \to \infty$. Assume I_{min} has a_n predictors where $1 \leq a_n \leq p$.

b) Suppose S is not a subset of I. Under the model $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$, $corr(r, r_I)$ will not converge to 1 as $n \to \infty$. Suppose that for large enough n, $[corr(r, r_I)]^2 \leq \gamma < 1$. Show that $C_p(I) \to \infty$ as $n \to \infty$.

R problems

```
regbootsim3(nruns=500)
$cicov
0.942 0.954 0.950 0.948 0.944 0.946 0.946 0.940 0.938 0.940
$avelen
0.398 0.399 0.397 0.399 2.448 2.448 2.448 2.448 2.448 2.450
$beta
[1] 1 1 0 0
$k
[1] 1
```

7.15. Use the *R* command for this problem, and put the output in *Word*. The output should be similar to that shown above. Consider the multiple linear regression model $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + e_i$ where $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The function `regbootsim3` bootstraps the regression model with the residual bootstrap. Note that $S = \{1, 2\}$ and $E = \{3, 4\}$. The first 4 numbers are the bootstrap shorth confidence intervals for β_i. The lengths of the CIs along with the proportion of times (coverage) the CI for β_i contained β_i are given. The CI lengths for the first 4 intervals should be near 0.392. With 500 runs, coverage in [0.92,0.98] suggests that the actual coverage is near the nominal coverage of 0.95. The next three numbers test $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$ where E corresponds to the last $p-k+1$ β_i. The prediction region method, hybrid method, and Bickel and Ren methods are used. Hence the fifth interval gives the length of the interval $[0, D_{(c)}]$ where H_0 is rejected if $D_0 > D_{(c)}$ and the fifth "coverage" is the proportion of times the prediction region method test fails to reject H_0. The last three numbers are similar but test $H_0 : \boldsymbol{\beta}_S = \boldsymbol{1}$ where S corresponds to the first $k+1$ β_i. Hence the last length 2.450 corresponds to the Bickel and Ren method with coverage 0.940. Want lengths near 2.45 which correspond to $\sqrt{\chi^2_2(0.95)}$ where $P(X \leq \chi^2_2(0.95)) = 0.95$ if $X \sim \chi^2_2$.