Homework 8

Due Thursday, 11/7/19

1. Variance inflation factors

   (a) Let $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ be a symmetric positive definite matrix (that means it's also invertible). Suppose $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_2^T & a_3 \end{bmatrix}$ where $\boldsymbol{A}_1 \in \mathbb{R}^{(p-1)\times(p-1)}$, $\boldsymbol{A}_2 \in \mathbb{R}^{p-1}$ and $a_3 \neq 0$ is a scalar. Show that

   $$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_1^{-1} + \tilde{a}^{-1}\boldsymbol{A}_1^{-1}\boldsymbol{A}_2\boldsymbol{A}_2^T\boldsymbol{A}_1^{-1} & -\tilde{a}^{-1}\boldsymbol{A}_1^{-1}\boldsymbol{A}_2 \\ -\tilde{a}^{-1}\boldsymbol{A}_2^T\boldsymbol{A}_1^{-1} & \tilde{a}^{-1} \end{bmatrix}, \quad \tilde{a} = a_3 - \boldsymbol{A}_2^T\boldsymbol{A}_1^{-1}\boldsymbol{A}_2.$$

   (b) Let $\boldsymbol{X} = \left[\boldsymbol{1}_n \, \boldsymbol{X}_1 \cdots \boldsymbol{X}_{p-1}\right] \in \mathbb{R}^{n \times p}$ be a full rank design matrix.

      (i) Use part (a) to show that the $[j+1]$th diagonal element of $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ is

      $$\left\{\boldsymbol{X}_j^T\left(\boldsymbol{I}_n - \boldsymbol{H}_{(-j)}\right)\boldsymbol{X}_j\right\}^{-1},$$

      where $\boldsymbol{H}_{(-j)}$ is the hat matrix for $\left[\boldsymbol{1}_n \, \boldsymbol{X}_1 \cdots \boldsymbol{X}_{j-1} \, \boldsymbol{X}_{j+1} \cdots \boldsymbol{X}_{p-1}\right]$.

      (ii) Will this change if we mean-center $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{p-1}$? Why or why not?

   (c) Show that the variance inflation factor (VIF) for the jth predictor, defined on the last slide of the lecture slides from 10/29, can be written as

   $$VIF_j = \frac{\boldsymbol{X}_j^T\left(\boldsymbol{I}_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^T\right)\boldsymbol{X}_j}{\boldsymbol{X}_j^T\left(\boldsymbol{I}_n - \boldsymbol{H}_{(-j)}\right)\boldsymbol{X}_j}.$$

   (d) Use part (c) to show that

      (i) $VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination for the regression of $\boldsymbol{X}_j$ onto $\boldsymbol{1}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_{p-1}$

      (ii) $VIF_j \geq 1$ and $VIF_j = \infty$ if and only if $R_j^2 = 1$.

      (iii) Using parts (i) and (ii), explain why a large variance inflation factor implies nearly all of the variance in $\boldsymbol{X}_j$ can be explained by the other predictors.

   (e) Now consider the data SeatPos.txt (the file SeatPosDescription.txt contains a description of the data).

      (i) Regress ”hipcenter” onto all covariates. What is the P value for the null hypothesis that the coefficients for all predictors are 0? Would you reject the null hypothesis? Based on the coefficient of determination and the hypothesis test results, would you conclude that these covariates do a good job at predicting hipcenter?

      (ii) From the regression output from part (i), are there any individual covariates that appear to be significantly related to hipcenter?

(iii) Now compute the variance inflation factors for each of the eight predictors. Are there any that appear to be aberrantly large?

(iv) Let $\hat{\beta}_{HS}$ be the coefficient for "HtShoes" from part (i). Generate 100 synthetic datasets by adding independently and identically distributed noise of the form $N(0,1)$ to the dependent variable hipcenter (i.e. you alter hipcenter by ONLY 1mm, on average). For each dataset $b = 1,\ldots,100$, re-compute the estimate for the coefficient for "HtShoes" (call it $\hat{\beta}_{HS}^{(b)}$) and report

$$\sqrt{100^{-1}\sum_{b=1}^{100}\left(\frac{\hat{\beta}_{HS}-\hat{\beta}_{HS}^{(b)}}{\hat{\beta}_{HS}}\right)^2}.$$

Given the variance inflation factor for "HtShoes", are you surprised that this is so large? Why does this help show that a large variance inflation factor implies the estimate for the corresponding coefficient is very uncertain?

(v) Now report the correlation matrix for the 8 predictors and show that HtShoes, Ht, Seaterd, Arm, Thigh and Leg are very correlated with one another. Are you surprised by this? (Hint: these are all related to one's height!)

(vi) Remove the predictors HtShoes, Seated, Arm, Thigh and Leg from the regression you fit in part (i). Explain why the coefficient of determination changes very little, but the P value for Ht is now very significant.