STA 712 Homework 3

Due: Friday, September 23, 12:00pm (noon) on Canvas.

Instructions: Submit your work as a single PDF. For this assignment, you may include written work by scanning it and incorporating it into the PDF. Include all R code needed to reproduce your results in your submission.

Variance Inflation Factors

In class, we introduced *variance inflation factors* as a method for diagnosing multicollinearity. In class, we said that the variance inflation factor VIF_j for the coefficient $\widehat{\beta}_j$ is given by

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination for the linear regression of the jth explanatory variable on the other explanatory variables. The goal of the problems in this section is to derive this variance inflation factor.

1. Before we can derive the variance inflation factor, we need to derive some properties of the coefficient of determination R^2.

   Ignore logistic regression for now, and suppose we have the *linear* regression model

   $$Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon,$$

   where $\boldsymbol{X} \in \mathbb{R}^{n \times (k+1)}$ is the design matrix and $\varepsilon \sim N(0, \sigma^2 I)$. The coefficient of determination R^2 for the regression of Y on $\boldsymbol{X}$ is given by

   $$R^2 = 1 - \frac{SSE}{SSTotal} = 1 - \frac{\sum_i (Y_i - \widehat{Y}_i)^2}{\sum_i (Y_i - \overline{Y})^2}.$$

   For the purposes of this question, assume that $\overline{Y} = 0$ (this will make our math easier).

   (a) Let $H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ be the hat matrix for this linear regression. Show that

   $$SSE = Y^T(I - H)Y.$$

   (b) Let $H_0 = \mathbf{1}(\mathbf{1}^T\mathbf{1})^T\mathbf{1}^T$, where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all 1s. Show that

   $$SSTotal = Y^T(1 - H_0)Y.$$

   (That is, the total sum of squares is just the residual sum of squares when we regress on a constant).

   (c) Using (a) and (b), and the assumption that $\overline{Y} = 0$, show that

   $$R^2 = \frac{Y^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{Y^TY}.$$

2. In this question, we will derive variance inflation factors for logistic regression.

We will work with the logistic regression model

$$Y_i \sim Bernoulli(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}.$$

Let $\mathbf{x}_i = (X_{1,i}, X_{2,i}, ..., X_{n,i})^T \in \mathbb{R}^n$ denote the vector of observed responses for the ith explanatory variable. Then, the design matrix for our logistic regression model can be written

$$\boldsymbol{X} = [\mathbf{1}\ \mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_k] \in \mathbb{R}^{n\times(k+1)},$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all 1s.

Letting $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^T \in \mathbb{R}^{k+1}$, recall from class that

$$Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1},$$

where $\boldsymbol{W}$ is the diagonal weight matrix from the last iteration of Fisher scoring, with diagonal entries $w_i = \widehat{p}_i(1-\widehat{p}_i)$.

For the purposes of this problem, assume that the columns $\mathbf{x}_i$ have been standardized so that $\sum\limits_{j=1}^{n} w_j X_{j,i} = 0$. This standardization does not impact the correlation between the columns, and therefore does not impact the variance inflation factors, but it makes our math easier.

At several points in this problem, it will be helpful to use the following fact about inverting block matrices. Let

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

be a block matrix with $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{p\times q}$, $C \in \mathbb{R}^{q\times p}$, and $D \in \mathbb{R}^{q\times q}$. Assuming that A and D are invertible, then

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}.$$

(a) Show that

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} = \begin{bmatrix} \sum\limits_{i=1}^{n} w_i & \mathbf{0}^T \\ \mathbf{0} & (\boldsymbol{X}^*)^T\boldsymbol{W}\boldsymbol{X}^* \end{bmatrix},$$

where $\boldsymbol{X}^* = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_k] \in \mathbb{R}^{n \times k}$ and $\mathbf{0} \in \mathbb{R}^k$. Conclude that

$$Var(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \dfrac{1}{\sum\limits_{i=1}^{n} w_i} & \mathbf{0}^T \\ \mathbf{0} & ((\boldsymbol{X}^*)^T\boldsymbol{W}\boldsymbol{X}^*)^{-1} \end{bmatrix}.$$

(b) Using (a), show that if our only explanatory variable is $\mathbf{x}_1$, that is, $\log\left(\dfrac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1}$, then $Var(\widehat{\beta}_1) = (\mathbf{x}_1^T\boldsymbol{W}\mathbf{x}_1)^{-1}$.

(c) Now we want to calculate $Var(\widehat{\beta}_1)$ when other explanatory variables are included in the model. Write

$$\boldsymbol{X}^* = [\mathbf{x}_1 \ \boldsymbol{X}^*_{(1)}],$$

where $\boldsymbol{X}^*_{(1)} = [\mathbf{x}_2 \ \cdots \ \mathbf{x}_k]$. Show that

$$(\boldsymbol{X}^*)^T\boldsymbol{W}\boldsymbol{X}^* = \begin{bmatrix} \mathbf{x}_1^T\boldsymbol{W}\mathbf{x}_1 & \mathbf{x}_1^T\boldsymbol{W}\boldsymbol{X}^*_{(1)} \\ (\boldsymbol{X}^*_{(1)})^T\boldsymbol{W}\mathbf{x}_1 & (\boldsymbol{X}^*_{(1)})^T\boldsymbol{W}\boldsymbol{X}^*_{(1)}. \end{bmatrix}$$

Conclude that when other explanatory variables are included in the model,

$$Var(\widehat{\beta}_1) = (\mathbf{x}_1^T\boldsymbol{W}\mathbf{x}_1 - \mathbf{x}_1^T\boldsymbol{W}\boldsymbol{X}^*_{(1)}((\boldsymbol{X}^*_{(1)})^T\boldsymbol{W}\boldsymbol{X}^*_{(1)})^{-1}(\boldsymbol{X}^*_{(1)})^T\boldsymbol{W}\mathbf{x}_1)^{-1}.$$

(d) Now consider a weighted least squares regression of $\mathbf{x}_1$ on the other $k-1$ explanatory variables $\mathbf{x}_2, ..., \mathbf{x}_k$, with weights $\boldsymbol{W}$. That is, we model

$$\begin{aligned} \mathbf{x}_1 &= \boldsymbol{X}^*_{(1)}\boldsymbol{\gamma} + \varepsilon \\ &= \gamma_1\mathbf{x}_2 + \gamma_2\mathbf{x}_3 + \cdots + \gamma_{k-1}\mathbf{x}_k + \varepsilon, \end{aligned}$$

with $\varepsilon \sim N(\mathbf{0}, \boldsymbol{W}^{-1})$. Use the derivation of the weighted least squares coefficient estimates from class to show that

$$\widehat{\boldsymbol{\gamma}} = ((\boldsymbol{X}^*_{(1)})^T\boldsymbol{W}\boldsymbol{X}^*_{(1)})^{-1}(\boldsymbol{X}^*_{(1)})^T\boldsymbol{W}\mathbf{x}_1,$$

and therefore

$$Var(\widehat{\beta}_1) = (\mathbf{x}_1^T\boldsymbol{W}\mathbf{x}_1 - \mathbf{x}_1^T\boldsymbol{W}\boldsymbol{X}^*_{(1)}\widehat{\gamma})^{-1}.$$

(e) Now we want to simplify this variance somewhat.