STA 712 Challenge Assignment 4: Deriving Variance Inflation Factors

Due: Wednesday, October 12, 12:00pm (noon) on Canvas.

Instructions:

- Submit your work as a single PDF. All work should be typed, with math and equations typeset using LaTeX or similar software.
- You are welcome to work with others on this assignment, but you must submit your own work.
- You can probably find the answers to many of these questions online. It is ok to use online resources! But make sure to show all your work in your final 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 this challenge assignment 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

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

(d) Now suppose we have the weighted linear regression model

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim N(0, W^{-1}),$$

where $W = \text{diag}(w_1, ..., w_n)$ is a diagonal matrix of weights. As we discussed in class, we can express this model as unweighted linear regression

$$Y_w = \boldsymbol{X}_w\boldsymbol{\beta} + \varepsilon_w, \qquad \varepsilon \sim N(0, I),$$

by transforming: $Y_w = W^{\frac{1}{2}}Y$, $\boldsymbol{X}_w = W^{\frac{1}{2}}\boldsymbol{X}$, and $\varepsilon_w = W^{\frac{1}{2}}\varepsilon$. Assume that $\overline{Y}_w$ is centered so that $\overline{Y}_w = \sum_{j=1}^n w_j^{\frac{1}{2}} Y_j = 0$ (note that centering Y_w does not change the estimated coefficients, except for the intercept β_0, which we usually don't care about). Use (c) to show that the coefficient of determination for the weighted least squares model is

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

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 with diagonal entries $w_i = p_i(1-p_i)$.

For the purposes of this problem, assume that the columns $\mathbf{x}_i$ have been centered so that $\sum_{j=1}^{n} w_j^{\frac{1}{2}} X_{j,i} = 0$. This centering does not impact the correlation between the columns, and therefore does not impact the variance inflation factors, but it makes some of 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) We will begin by finding an expression for $Var(\widehat{\beta}_1)$ (the argument is analogous for the other $\widehat{\beta}_j$). First, note that the ordering of the columns of $\boldsymbol{X}$ doesn't change our estimated regression coefficients, just the order in which they appear in the vector $\widehat{\boldsymbol{\beta}}$. Since we want to focus on $\widehat{\beta}_1$, let $\boldsymbol{X}^*$ be the reordered columns of $\boldsymbol{X}$, with

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

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

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

(b) 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

$$\begin{aligned} \mathbf{x}_1 &= \boldsymbol{X}^*_{(1)} \boldsymbol{\gamma} + \varepsilon \\ &= \gamma_1 \mathbf{1} + \gamma_2 \mathbf{x}_2 + \gamma_3 \mathbf{x}_3 + \cdots + \gamma_k \mathbf{x}_k + \varepsilon, \end{aligned}$$

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

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

(c) Now we want to simplify this variance somewhat. Using the results from part (b) and Question 1(d), show that

$$Var(\widehat{\beta}_1) = \frac{1}{(1 - R_1^2) \mathbf{x}_1^T \boldsymbol{W} \mathbf{x}_1},$$

where R_1^2 is the coefficient of determination for the weighted least squares regression of $\mathbf{x}_1$ on $\boldsymbol{X}^*_{(1)}$, with weights $\boldsymbol{W}$.

(d) Now we need to determine what $Var(\widehat{\beta}_1)$ would be if $\mathbf{x}_1$ were the *only* explanatory variable in the model. First, re-center $\mathbf{x}_1$ so that $\sum_{j=1}^{n} w_j X_{j,1} = 0$. Then suppose our model is

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{i,1}.$$

(Note that re-centering $\mathbf{x}_1$ does not change the slope β_1). Show that

$$[\mathbf{1} \ \mathbf{x}_1]^T \boldsymbol{W} [\mathbf{1} \ \mathbf{x}_1] = \begin{bmatrix} \sum_{i=1}^{n} w_i & 0 \\ 0 & \mathbf{x}_1^T \boldsymbol{W} \mathbf{x}_1 \end{bmatrix}.$$

Conclude that if $\mathbf{x}_1$ is our only explanatory variable in the model, then $Var(\widehat{\beta}_1) = (\mathbf{x}_1^T \boldsymbol{W} \mathbf{x}_1)^{-1}$.

(e) By comparing (c) and (d), show that when other explanatory variables $\mathbf{x}_2, ..., \mathbf{x}_k$ are added to the model, the variance of $\widehat{\beta}_1$ increases by a factor $\dfrac{1}{1 - R_1^2}$.