Kevin Su

28 Oct 2021

STA 206 Homework 4

1. Derive $E[SST]$ and $E[SSR]$ under the SLR model using matrix algebra.

$SST = \sum (Y_i - \bar{Y})^2$ $\quad \bar{Y} = \frac{1}{n}(J_n)Y$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \bar{Y} = \frac{1}{n} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$I_n Y - \frac{1}{n}(J_n)Y = Y - \frac{1}{n}(J_n)Y = \begin{bmatrix} (Y_1 - \bar{Y}) \\ (Y_2 - \bar{Y}) \\ \vdots \\ (Y_n - \bar{Y}) \end{bmatrix}$$

$$E[SST] = \left[\left(I_n - \frac{1}{n}(J_n)\right)Y\right]^T \left[\left(I_n - \frac{1}{n}(J_n)\right)Y\right]$$

$$= Y^T \underbrace{\left(I_n - \frac{1}{n}(J_n)\right)^T \left(I_n - \frac{1}{n}(J_n)\right)}_{\text{projection matrix}} Y$$

$$= Y^T \left(I_n - \frac{1}{n}(J_n)\right) Y$$

1 (Con't)

$E[SSE] = E[(\hat{Y} - \bar{Y})^2]$

$\hat{Y} = HY \qquad \bar{Y} = \frac{1}{n} J_n Y$

$$\begin{bmatrix} \hat{Y}_1 - \bar{Y} \\ \hat{Y}_2 - \bar{Y} \\ \vdots \\ \hat{Y}_n - \bar{Y} \end{bmatrix}$$

$$E\left[(\hat{Y} - \bar{Y})^2\right] = \left(HY - \frac{1}{n} J_n Y\right)^T \left(HY - \frac{1}{n} J_n Y\right)$$

$$= \left(\left(H - \frac{1}{n} J_n\right) Y\right)^T \left(\left(H - \frac{1}{n} J_n\right) Y\right)$$

$$= Y^T \underbrace{\left[H - \frac{1}{n} J_n\right]^T \left[H - \frac{1}{n} J_n\right]}_{\text{Proj Matrix}} Y$$

$$= Y^T \left[H - \frac{1}{n} J_n\right] Y$$

3. For each of the following models, answer whether it can be expressed as a multiple regression model or not. If so, indicate which transformations and/or new variables need to be introduced.

(a) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$

No transformations/new variables needed, already expressed as a multiple regression

(b) $Y_i = \varepsilon_i \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)$. $(\varepsilon_i > 0)$

Transform Y_i's $\longrightarrow \log(\hat{Y}_i)$ such that

$\log(\hat{Y}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}^2 + \varepsilon_i$

(c) $Y_i = \beta_0 \exp(\beta_1 X_{i1}) + \varepsilon_i$

Cannot be expressed as a multiple regression model.

(d) $Y_i = \{1 + \exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)\}^{-1}$

$\ln\left(\frac{1}{Y_i} - 1\right) = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$

↑

Transform response to $\ln\left(\frac{1}{Y_i} - 1\right)$

4. (a) What is the maximum number of X variables that can be included in a multiple regression model with intercept that is used to fit a data set with 10 cases?

There is no maximum number of X variables that can be included, but at $n-1 = 9$ X variables, the model would go through every point and this overfitted model will have no value in interpretability.

(b) With 4 predictors, how many X variables are there in the interaction model with all main effects and all interaction terms (2nd order / 3rd order, etc)?

There are $\binom{4}{4} + \binom{4}{3} + \binom{4}{2} + \binom{4}{1} + \binom{4}{0} = 16$ X variables in the interaction model with all main effects and interaction terms.

5. T/F

(a) The multiple coefficient of determination R^2 is always larger/not smaller for models with more X variables.

False. This is only true if the set of variables in the smaller model is a subset of the set of variables in the larger model.

(b) If all the regression coefficients associated with the X variables are estimated to be zero, then $R^2 = 0$.

True in general. However, if $SST = 0$, R^2 = indeterminate?

5 (c) The adjusted multiple coefficient of determination R^2_{adj} may decrease when adding additional X variables into the model.

True. If the additional gain in R^2 does not outweigh the additional penalty from adding another parameter, then R^2_{adj} will decrease.

(d) Models with larger R^2 is always preferred.

False, R^2_{adj}, AIC, BIC, Mallows C_p are better criterions for model selection.

(e) If the response vector is a linear combination of the columns of the design matrix X, then the coefficient of multiple determination $R^2 = 1$.

True. This is an example of overfitting in the extreme case.

7. Under the multiple regression model, show that the residuals are uncorrelated with the fitted values and the estimated regression coefficients.

$\Rightarrow$ Show $Cov(\hat{Y}, e) = 0$

$\hat{Y} = HY \qquad e = Y - HY$

$$Cov(HY, Y - HY) = Cov(HY, Y) - Cov(HY, HY)$$
$$H\,Var(Y) - H\,Var(Y) = 0$$

H a projection matrix