STAT 2509B4
Assignment 4

Krystian Wojcicki, 101001444

Winter 2020

Christmas week is a critical period for most ski resorts. Because many students and adults are free, they are able to spend several days indulging in their favorite pastime, skiing. A ski resort in Vermont wanted to determine the effect that weather had on their sales of lift tickets. The manager of the resort collected the number of lift tickets sold during the Christmas week (y), the total snowfall (x1) and the average temperature (x2) for the past 20 years (x3). The TSS for the full model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ is : TSS = 56601012

We decided to screen the independent variables to determine the best set for predicting the lift tickets sales. The sums of squares for all possible regression models were found to be as follows:

(a) **Determine the subset of variables that is selected as best by the Forward Selection Procedure using $F_0* = 4.2$ (to-add-variable). Show your steps.**

(1) Fit all one term models: $y = \beta_0 + \beta_1 x_j + \epsilon$ for $j = 1, 2, 3$
$\text{SSR}(X_1) = 6215561$
$\text{SSR}(X_2) = 665121$
$\text{SSR}(X_3) = 35788320 \Rightarrow \max$

Therefore $F_3 = \frac{\text{MSR}(X_3)}{\text{MSE}(X_3)} = \frac{\text{SSR}(X_3)/1}{\text{SSE}(X_3)/18} = \frac{35788320/1}{20812692/18} = 30.952$
Since $F_3 = 30.952 > F_0 = 4.2$ we keep X_3

(2) Fit all two term models $y = \beta_0 + \beta_1 x_3 + \beta_2 x_j + \epsilon$ for $j = 1, 2$

Calculate $\text{SSR}(X_j|X_3)$

$\text{SSR}(X_1|X_3) = \text{SSR}(X_1, X_3) - \text{SSR}(X_3) = 41296990 - 35788320 = 5508670 \Rightarrow \max$
$\text{SSR}(X_2|X_3) = \text{SSR}(X_2, X_3) - \text{SSR}(X_3) = 36518115 - 35788320 = 729795$
Therefore $F_1 = \frac{\text{MSR}(X_1|X_3)}{\text{MSE}(X_1,X_3)} = \frac{[\text{SSR}(X_1,X_3) - \text{SSR}(X_3)]/[df_{SSR(X_1,X_3)} - df_{SSR(X_3)}]}{\text{SSE}(X_1,X_3)/df_{SSE(X_1,X_3)}} = \frac{5508670/(2-1)}{15304022/17} = 6.1191358716$
Since $F_1 = 6.1191358716 > 4.2$ we keep X_1, X_3

(3) Fit the full model $y = \beta_0 + \beta_1 x_3 + \beta_2 x_1 + \beta_3 x_3 + \epsilon$
Calculate $\text{SSR}(X_2|X_1, X_3) = \text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1, X_3) = 41940217 - 41296990 = 643227$
Therefore $F_2 = \frac{\text{MSR}(X_2|X_1,X_3)}{\text{MSE}(X_1,X_2,X_3)} = \frac{[\text{SSR}(X_1,X_2,X_3) - \text{SSR}(X_1,X_3)]/[df_{SSR(X_1,X_2,X_3)} - df_{SSR(X_1,X_3)}]}{\text{SSE}(X_1,X_2,X_3)/df_{SSE(X_1,X_2,X_3)}} = \frac{643227/(3-2)}{14660795/16} = 0.701983214416$
Since $F_2 \leq F_0$ we keept X_1, X_3

Therefore the best set is X_1, X_3

(b) **Determine the subset of variables that is selected as best by the Backward Elimination Procedure using $F_0 ** = 4.1$ (to-delete-variable). Show your steps. NOTE: $(t0^{**})^2 = F_0^{**}$**

(c) **Determine the subset of variables that is selected as best by the Stepwise Regression Procedure using F0 * = 4.2 (to-add) and F0 ** = 4.1 (to-delete). Show your steps.**

Fit the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ and check whether model is significant or not at $\alpha = 5\%$.

$F = \frac{\text{MSR}_f}{\text{MSE}_f} = \frac{\text{SSR}_f/3}{\text{SSE}_f/16} = \frac{41940217/3}{14660795/16} = 15.2570960397$

Since $F = 15.2570960397 > F_{(3,16);0.05} = 3.24$, we can conclude that at 5% level of sigificance the full model is significant and can be used

(1) Calculate $F_j = (t_j)^2 = \frac{\text{MSR}(X_j|\text{all}X\text{'s except}X_j)}{\text{MSE}(X_1,X_2,X_3)} = \frac{[SSR_f - SSR(\text{all}X\text{'s except}X_j)]/df}{MSE_f}$