Multiple Regression Examples

Full Model: \( y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i \)

This example is based on the example given in Introduction to the Practice of Statistics, Second Edition by David S. Moore and George P. McCabe on pages 686-704 and the data set CSDATA described on pages 792 and 793. The program used to generate the output is programs/csdata.sps.

**Question 1:** Is the grade point average linearly related to at least one of the 5 proposed explanatory variables? (i.e., does the full model "fit" the data?)

**H_0:** The grade point average of a computer science major at the end of three semesters is not linearly related to any of the variables: high school math grade (HSM), high school English grade (HSE), high school science grade (HSS), SAT math score (SATM), or SAT verbal score (SATV). (\( \beta_j = 0 \) for at least one \( j = 1, 2, \ldots, p \))

**H_a:** The grade point average of a computer science major at the end of three semesters is linearly related to at least one of the variables: high school math grade (HSM), high school English grade (HSE), high school science grade (HSS), SAT math score (SATM), or SAT verbal score (SATV). (\( \beta_j \neq 0 \) for at least one \( j = 1, 2, \ldots, p \))

Assumptions: The mean response is a linear function of the explanatory variables. The deviations are independent and normally distributed with mean 0 and standard deviation \( \sigma \) (i.e. \( \text{N}(0, \sigma^2) \)).

Test Statistic: \( F = \frac{\text{MSM}}{\text{MSE}} \), \( df_M = p = 5 \) and \( df_E = n - p - 1 = 218 \)

\( \alpha = .05 \)

RR: \( F > F_{.05,p,n-p-1} = F_{.05,5,218} \approx F_{.05,5,200} = 2.26 \)

or p-value of \( F < .05 \)

Calculation: From SPSS \( F = 11.69138 \) with p-value .0000

Decision: Reject \( H_0 \).

Conclusion: There is sufficient evidence to indicate that the grade point average of a computer science major at the end of three semesters is linearly related to at least one of the variables: high school math grade (HSM), high school English grade (HSE), high school science grade (HSS), SAT math score (SATM), or SAT verbal score (SATV).

Since we answered yes to question 1, there are two other types of questions that we might want to ask.
Multiple Regression Examples

**First:** Given that the full model fits the data do we need to include the variable, \( X_j \)?

\( H_a \): Given that the variables high school math grade (HSM), high school English grade (HSE), SAT math score (SATM), and SAT verbal score (SATV) are in the regression equation for the grade point average of a computer science major at the end of three semesters, the variable high school science grade (HSS) is needed in the model. \( (\beta_{\text{HSS}} \neq 0) \)

\( H_o \): Given that the variables high school math grade (HSM), high school English grade (HSE), SAT math score (SATM), and SAT verbal score (SATV) are in the regression equation for the grade point average of a computer science major at the end of three semesters, the variable high school science grade (HSS) is not needed in the model. \( (\beta_{\text{HSS}} = 0) \)

Assumptions: See previous assumptions.

Test Statistic: \( t = \frac{b_j}{s_{b_j}} \), \( df = n-p-1 = 218 \)

\( \alpha = .05 \) RR: \( t < -1.984 \) or \( t > 1.984 \) or \( p\)-value of \( t \leq .05 \)

Calculation: From SPSS \( t = .950 \) with \( p\)-value = .3432

Decision: Fail to reject \( H_o \).

Conclusion: There is not sufficient evidence to indicate that the variable high school science grade (HSS) is needed in the model, given that the variables high school math grade (HSM), high school English grade (HSE), SAT math score (SATM), and SAT verbal score (SATV) are in the regression equation for the grade point average of a computer science major at the end of three semesters.
Multiple Regression Examples

Second: Given that the full model fits the data do we need to include the variables, \(X_q, X_{q+1}, \ldots, X_p\)?

\(H_a:\) Given that the variables SAT math score (SATM), and SAT verbal score (SATV) are in the regression equation for the grade point average of a computer science major at the end of three semesters, at least one of the variables high school science grade (HSS), high school math grade (HSM) and high school English grade (HSE) is needed in the model. (\(\beta_j \neq 0\) for some \(j = \text{HSS, HSE or HSM}\))

\(H_o:\) Given that the variables SAT math score (SATM), and SAT verbal score (SATV) are in the regression equation for the grade point average of a computer science major at the end of three semesters, the variables high school science grade (HSS), high school math grade (HSM) and high school English grade (HSE) are not needed in the model. (\(\beta_{\text{HSS}} = \beta_{\text{HSE}} = \beta_{\text{HSM}} = 0\) for \(j = \text{HSS, HSE and HSM}\))

Assumptions: See assumptions for first test.

Test Statistic: 
\[
F = \frac{\text{SSE(Reduced)} - \text{SSE(Full)}}{\text{df}_{\text{SSSEReduced}} - \text{df}_{\text{SSSEFull}}} / \frac{\text{MSE(Full)}}{\text{df}_{\text{SSSEReduced}} - \text{df}_{\text{SSSEFull}}} \cdot \frac{\text{df}_{\text{num}} - \text{df}_{\text{SSSEReduced}} - \text{df}_{\text{SSSEFull}}}{\text{df}_{\text{den}} - \text{df}_{\text{SSSEFull}}}
\]

\(\alpha = .05\) \(RR: F > F_{.05, 3, 218} = F_{.05, 3, 200} = 2.65\) or p-value of \(F \leq .05\)

\[
F = \frac{\text{SSE(Reduced)} - \text{SSE(Full)}}{\text{MSE(Full)}}
\]

Calculation:

\[
\frac{126.87895 - 106.81914}{221 - 218} = 13.65
\]

p-value of \(F < .001\). The values used in this calculation are from the SPSS output for the full and reduced models.

Decision: Reject \(H_o\).

Conclusion: There is sufficient evidence to indicate that at least one of the variables high school science grade (HSS), high school math grade (HSM) and high school English grade (HSE) is needed in the model, given that the variables SAT math score (SATM), and SAT verbal score (SATV) are in the regression equation for the grade point average of a computer science major at the end of three semesters.