MULTICOLINEARITY PROBLEM
- Correlation Analysis
- Variance Inflation Factors

STEPWISE REGRESSION
- Mixed Method
- P-Values versus R-Square

ADDITIONAL DIAGNOSTICS
- Heteroscedasticity
- Autocorrelation
STUDENT PERFORMANCE
PROBLEM

It is sometimes argued by students (and even professors) that **final exams are unnecessary** because they have little effect in changing students’ grades. This claim can be tested by constructing a regression which predicts final-exam performance from midterm-exam performance.

**DATA:** Grading records have been obtained for a representative course with 29 students where a single midterm and final exam were given. This data includes the **final exam score**, $F_i$, **midterm exam score**, $M_i$, and **average homework score**, $H_i$, for each student $i = 1,..,29$.

**ANALYSIS:** Consider the following two regressions:

\[ F_i = \beta_0 + \beta_1 M_i + \varepsilon_i , \quad i = 1,..,n \]

\[ F_i = \beta_0 + \beta_1 M_i + \beta_2 H_i + \varepsilon_i , \quad i = 1,..,n \]
MULTICOLINEARITY PROBLEM

• PERFECT MULTICOLINEARITY

\[ x_2 = a + b x_1 \]

• COEFFICIENT INSTABILITY
VARIANCE INFLATION FACTOR

DEFINITION

Given a multiple regression model

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \]

let \( R_j^2 \) denote the R-square value for the regression of \( x_j \) on all other explanatory variables, \( (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k) \). The variance inflation factor (VIF\(_j\)) for variable \( x_j \) is then defined to be

\[ VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, k \]

INTERPRETATION

If \( \hat{\beta}_j \) is the regression estimate of \( \beta_j \) in (1), then it can be shown that

\[ \text{var}(\hat{\beta}_j) = VIF_j \left( \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2} \right) \]

But the expression in brackets is precisely the variance of \( \hat{\beta}_j \) in the simple regression model

\[ y_i = \beta_0 + \beta_j x_{ji} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n \]

without the other explanatory variable. So VIF\(_j\) reflects the change in variance of \( \hat{\beta}_j \) due to the presence of the other variables, i.e., the variance inflation factor for variable \( x_j \).