Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Model Evaluation and Multicollinearity

1

About This Lesson

2

R^2: Coefficient of Determination

A measure that efficiently summarizes how well the *X*s can be used to predict *Y* is $\mathbf{R^2}$ (called *R-squared* or the *coefficient of determination*):

$$R^2 = 1 - SSE/SST$$

where

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R^2 is interpreted as the proportion of total variability in *Y* that can be explained by the linear regression model.

Georgia Tech

R^2: Notation and Terminology

SSE, **SST**, and **SSR** refer to ***sum of squared errors***, ***sum of squares total***, and ***sum of squares for regression***. Unfortunately, the field of statistics abounds in inconsistent terminology and notation.

- **SSE:** sometimes denoted $\mathbf{SS_{error}}$ or $\mathbf{SS_{err}}$, is also known as **RSS** (residual sum of squares) and $\mathbf{SS_{res}}$ (sum of squared residuals, sometimes **SSR**).
- **SST:** sometimes written as $\mathbf{SS_{total}}$ or $\mathbf{SS_{tot}}$. It is also called total sum of squares and written as **TSS**.
- **SSR:** also called the sum of squares due to regression, and it is sometimes written as $\mathbf{SS_{reg}}$. It's also called explained sum of squares (**ESS**). Don't confuse ESS with SSE, and, for R^2, remember that SSR is SS regression, not SS residuals!

Georgia Tech

Model Evaluation

- **F-test for overall regression**
  - $\mathrm{H}_0: \beta_1 = \beta_1 = \cdots = \beta_p = 0$
  - $\mathrm{F}_0 = \mathrm{MSR}/\mathrm{MSE} \sim \mathrm{F}(p, n\text{-}p\text{-}1)$
    - $\mathrm{MSR} = \mathrm{SSR}/p$
      - $\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
    - $\mathrm{MSE} = \mathrm{SSE}/(n\text{-}p\text{-}1)$
- **Coefficient of determination**
  - $\mathrm{R}^2 = 1 - \mathrm{SSE}/\mathrm{SST} = \mathrm{SSR}/\mathrm{SST}$
  - R^2 increases when additional predictors are added to a model
    - Such increase might not indicate increased explanatory power
- **Adjusted coefficient of determination**
  - Penalizes for more predictors
  - adjusted $\mathrm{R}^2 = 1 - (n\text{-}1)(1\text{-}\mathrm{R}^2)/(n\text{-}p\text{-}1)$

Georgia Tech

5

Correlation Coefficient

A statistic that efficiently summarizes how well one of the *X*s is *linearly* related to *Y* (or to another *X*) is ρ, the (Pearson) correlation coefficient:

$$\rho_{X_j,Y} = \mathrm{cor}(X_j, Y) = \frac{\sum_{i=1}^{n}(x_{i,j} - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Can be used to evaluate the *linear* relationship between the response variable and any of the predicting variables, X_j
  - Useful when looking for transformations of predicting variables
- Can also be used to evaluate correlation between predicting variables
  - Can help detect near linear dependence (multicollinearity)

Georgia Tech

6

Multicollinearity

Recall that finding the ordinary least squares estimator of $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

depends on $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ being invertible (nonsingular or nondegenerate). From linear algebra, a square matrix is invertible if and only if its columns are linearly independent (i.e., no column is a linear combination of the others).

If that doesn't hold, the ordinary least squares estimator of $\widehat{\boldsymbol{\beta}}$ doesn't exist. That's probably due to a specification error where one or more predictors should be eliminated as redundant (e.g., if years and number of rings were included in a model for trees).

Even if the columns of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ are linearly independent, some problems might arise if the value of one predictor can be closely estimated from the other predictors. We call this condition *multicollinearity* or *near collinearity*.

Georgia Tech

7

Multicollinearity

- Indications that near collinearity is present:
  - The estimated coefficients $\widehat{\boldsymbol{\beta}}$ are unstable: When the value of one predictor changes slightly, the fitted regression coefficients change dramatically
  - The standard error of $\widehat{\boldsymbol{\beta}}$ is artificially large
  - The overall F statistic is significant, but individual *t*-statistics are not
- Prediction may be affected
  - The relationship to the response may change widely
- Some computational algorithms are sensitive to multicollinearity
- <u>But no</u> inflation or deflation in $\mathbf{R}^2$

Georgia Tech

8

Multicollinearity Diagnosis

Compute the variance inflation factor (VIF_j) for each predicting variable X_j

$$\boxed{\mathrm{VIF}_j = \frac{1}{1 - \mathrm{R}_j^2}}$$

where R_j^2 is the coefficient of determination for the regression of X_j against all other predicting variables.

What is an acceptable VIF (i.e., multicollinearity is not problematic)?

- $\mathrm{VIF} < \max\left(10,\ 1/\left(1 - \mathrm{R}^2_{\mathrm{model}}\right)\right)$
  - $\mathrm{R}^2_{\mathrm{model}}$ is the coefficient of determination for the original model
  - Rule of thumb only

Georgia Tech

9

Multicollinearity Diagnosis

Steps:

1. For $j = 1,\ldots,p$, regress X_j against all other X_i, $i = 1,\ldots,p,\ i \neq j$ (i.e., $X_j \sim X_1,\ldots,X_{j-1}, X_{j+1}, \ldots, X_p$).
2. For each regression run, compute R^2 for that regression (i.e., compute R_j^2 for $j = 1, \ldots, p$).
3. For each regression run, compute VIF_j based on the computed R_j^2 for that regression.
4. If any $\mathrm{VIF}_j \geq 10$ and it is also $\geq 1/\left(1 - \mathrm{R}^2_{\mathrm{model}}\right)$, where $\mathrm{R}^2_{\mathrm{model}}$ is the coefficient of determination for the original model, the test is positive for multicollinearity.

High multicollinearity is not detected if each $\mathrm{VIF}_j < \max\left(10, \frac{1}{1-\mathrm{R}^2_{\mathrm{model}}}\right)$.

Georgia Tech

10

Multicollinearity Interpretation

VIF measures the proportional increase in the variance of $\hat{\beta}_i$ compared to what it would have been if the predicting variables had been completely uncorrelated.

- VIF of 1 (the minimum possible VIF) means the tested predictor is not correlated with the other predictors
- The higher the VIF:
  - The more correlated a predictor is with the other predictors
  - The more the standard error is inflated
  - The larger the confidence interval
  - The less likely it is that a coefficient will be evaluated as statistically significant

Georgia Tech

11

Summary

Georgia Tech

12