Multiple Linear Regression

October 8, 2017

Multiple Linear Regression

Linear in parameters

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

- $X_{1i}, X_{2i}, \ldots, X_{pi}$ are p covariates measured on individual i
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are parameters (unknown)
- Y is the response variable
- ϵ_i $i = 1, \ldots, n$ are iid with mean 0 and variance σ^2
- ϵ_i often assumed $N(0, 1)$

Multiple Linear Regression

Includes models that are polynomial in a single covariate X

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots \beta_p X^p$$

Matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Least squares estimates

$$\mathbf{Q} = (\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta) = \boldsymbol{Y}'\boldsymbol{Y} - \beta'\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\beta + \beta'\boldsymbol{X}'\boldsymbol{X}\beta$$

$$\mathbf{Q} = \boldsymbol{Y}'\boldsymbol{Y} - 2\beta'\boldsymbol{X}'\boldsymbol{Y} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta$$

$$\frac{\partial \mathbf{Q}}{\partial \beta} = -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\beta$$

Set equal to zero:

$$\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\mathbf{b}$$

$$\mathbf{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Distributions

$$\boldsymbol{Y} \sim MVN(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I})$$

$$E(\mathbf{b}) = E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta = \beta$$

$$Var(\mathbf{b}) = var((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'var(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} =$$

$$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\mathbf{b} \sim MVN\left(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

Estimating σ^2

We can estimate σ^2 by

$$S^2 = \frac{SSE}{n - (p + 1)}$$

where $p + 1$ is the number of parameters estimated

ANOVA Table

Source	SS	df	MS	F
Regression	SSR	p	SSR/p	MSR/MSE
Error	SSE	n-(p+1)	SSE/(n-(p+1))	
Total	SST	n-1		

Example

- What influences a baby's birth weight? Suppose we have data on birth weight (response) and predictors based on parent's age, weight, height, gestation period and mother's activities during pregnancy.
- In 1609 Galileo proved the distance traveled by an object with an initial height is a parabola. He could have based this on experiments. Let Y be the distance travelled and X be the initial height. Could have considered these models.
  - $Y = \beta_0 + \beta_1 X$
  - $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ True model
  - $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

  All these models could fit the data equally well

Predicted Values and Residuals

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\mathbf{b} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \mathbf{H}\boldsymbol{Y}$$

where $\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

$$\mathbf{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbf{H}\boldsymbol{Y} = (\boldsymbol{I} - \mathbf{H})\boldsymbol{Y}$$

Residual Analysis

As before we use the residuals to check the fit of the model.

- Plot **e** vs each independent variable X_i to check whether higher order terms are needed
- Plot **e** vs $\hat{\boldsymbol{Y}}$ to check the assumption of constant variance
- Normal plot of **e** to check for normality
- plot **e** versus time order to look for serial correlation

Multicollinearity occurs when the columns of $\boldsymbol{X}$ are linearly dependent or nearly so. Practically this means that some of the independent variables are measuring the same thing and are not needed.

If the columns of $\boldsymbol{X}$ are linearly dependent then $\boldsymbol{X}'\boldsymbol{X}$ is not of full rank and is singular so $\mathbf{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ can't be computed. If the columns are nearly linearly dependent, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ has very large elements and since $Var(\mathbf{b}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, tests of $H_0 : \beta_i = 0$ will not be statistically significant.

Can check for multicollinearity by looking at correlations between predictor variables.

Use dummy Variables.

1. If we have a nominal variable with $c \geq 2$ categories, define $X_1, X_2, \ldots X_{c-1}$ dummy variables

$$X_i = \begin{cases} 1 & \text{for the } i\text{th category} \\ 0 & \text{otherwise} \end{cases}$$

If we define c dummy variables we will induce multicollinearity.
Example Male or Female can be coded with one dummy variable

$$X = \begin{cases} 1 & \text{for male} \\ 0 & \text{for female} \end{cases}$$

If we tried to use two variables

$$X_1 = \begin{cases} 1 & \text{for male} \\ 0 & \text{for female} \end{cases} \quad X_2 = \begin{cases} 1 & \text{for female} \\ 0 & \text{for male} \end{cases}$$

Then $\boldsymbol{X}_1 + \boldsymbol{X}_2 = \mathbf{1}$ which conflicts with the β_0 column in $\boldsymbol{X}$.

2 If ordinal variables, such as prognosis of a patient (poor, average, good), the categories can be assigned numerical scores (1,2,3) and be treated as numerical variables.

Obviously for p covariates there are many possible models. How do we select the best one? Partial F Test

- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$ is a partial model
- $Y = \beta_1 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ is the full model

The partial model has $k+1$ parameters and the full model has $p+1$ parameters $p > k$. The residual sum of squares, SSE, measures the variation between data and the model. Thus SSE for the full model can only be less than SSE for the partial model. SSE(partial) - SSE(full) is the extra sum of squares. If the new parameters $\beta_{k+1}, \ldots, \beta_p$ are not really important then the extra sum of squares should be smmall.

To test $H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$ compute statistic

$$F = \frac{(SSE(partial) - SSE(full))/(p-k)}{SSE(full)/(n-(p+1))}$$

Under $H_0, F \sim F_{p-k,n-(p+1)}$. Reject H_0 if F is too big.

- coefficient of determination, r^2, higher is better
- AIC - Akaike information criterion, lower is better
- BIC - Bayesian information criterion, lower is better