STAT 480 Statistical Computing Applications

Unit 3. Linear Regression

Lecture 2. Multiple Linear Regression

Department of Statistics, Iowa State University

Spring 2019

Multiple Linear Regression Model

- Multiple Linear Regression Models

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

for $i = 1, \ldots, n$.

  - X_{ij} represents the jth predictor for the ith subject;
  - β_j quantifies the association between that variable and the response;
  - We interpret β_j as the average effect on Y of a one unit increase in the jth predictor, holding all other predictors fixed.

Car Example

- Suppose that Y is the fuel consumption of a particular model of car in m.p.g. Suppose that the predictors are
  - X_1 – the weight of the car
  - X_2 – the horse power
  - X_3 – the number of cylinders.
- Typically the data will be available in the form of an array:

$$\begin{array}{cccc} y_1 & x_{11} & x_{12} & x_{13} \\ y_2 & x_{21} & x_{22} & x_{23} \\ \ldots & & \ldots & \\ y_n & x_{n1} & x_{n2} & x_{n3} \end{array}$$

  where n is the number of observations in the dataset, $p = 3$ is the number of predictors.

Notations

- Let

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix}, \quad \cdots, \quad \mathbf{X}_p = \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix},$$

- We collect the predictors into columns of a matrix $\mathbf{X}$.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times (p+1)}.$$

Multiple Linear Regression

- The model is:

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \cdots + \beta_p \mathbf{X}_p + \boldsymbol{\varepsilon}$$

- Alternatively, we can write the model as

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

  - $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the true coefficients,
  - the errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ are as before (i.e., satisfying $E(\boldsymbol{\varepsilon}) = 0$ and $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$).

Interpreting Regression Coefficients

- Then ideal scenario is when the predictors are uncorrelated – a balanced design:
  - Each coefficient can be estimated and tested separately.
  - Interpretation such as "a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed", are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically.
  - Interpretations become hazardous – when X_j changes, everything else changes.
- Claims of causality should be avoided for observational data!!

Estimation and Prediction

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p.$$

- We estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ as the values that minimize the sum of squared residuals

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

Estimation by Least Squares

- This gives

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

  (Check: does this match the expressions for univariate regression, without and with an intercept?)

- The fitted values are

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- This is a linear function of $\mathbf{Y}$, $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

  is sometimes called the hat matrix.

Remarks

- The number of observations (n) must be at least as large as the number of parameters ($p+1$); preferably much larger.
- If two (or more) variables X_{ij} and X_{ik} are exactly linearly dependent upon one another, the $(\mathbf{X}^T\mathbf{X})^{-1}$ will not be unique solution exists. Even if two variables are not exactly linear dependent, if they are highly correlated, attempting to put both in the model can cause strange estimates to arise
- If some X_j's are deleted from the model, one CANNOT simply delete the corresponding β_j's from the model and use the remaining coefficients. Rather, one must re-fit the regression using only the desired variables to obtain new estimates of the coefficients.

Life-Circle Saving

- Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by
  - per capita disposable income,
  - the percentage rate of change in per-capita disposable income
  - the percentage of population less than 15 years old
  - the percentage of the population over 75 years old.
- The data is a built-in R data named `LifeCycleSavings`.

Life-Circle Saving (Cont.)

The data frame contains 50 observations on 5 variables.

sr	aggregate personal saving
pop15	% of population under 15
pop75	% of population over 75
dpi	real per-capita disposable income
ddpi	% growth rate of dpi

	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
Bolivia	5.75	41.89	1.67	189.13	0.22
Brazil	12.88	42.19	0.83	728.47	4.56
Canda	8.79	31.72	2.85	1982.88	2.43

Life-Circle Saving (Cont.)

Correlations

	sr	pop15	pop75	dpi	ddpi
sr	1.000	-0.456	0.317	0.220	0.305
pop15	-0.456	1.000	-0.908	-0.756	-0.048
pop75	0.317	-0.908	1.000	0.787	0.025
dpi	0.220	-0.756	0.787	1.000	-0.129
ddpi	0.305	-0.048	0.025	-0.129	1.000

Coefficients

	Estimate	Std. Error	t value	p–value
Intercept	28.5661	7.355	3.884	0.0003
pop15	-0.4612	0.145	-3.189	0.0026
pop75	-1.6915	1.084	-1.561	0.1255
dpi	-0.0003	0.001	-0.362	0.7192
ddpi	0.4097	0.196	2.088	0.0425

Some Important Questions

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is At Least One Predictor Useful?

- Is there a relationship between the response and predictors?
- We test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.

ANOVA Table for Linear Model

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	p	$\sum (\hat{y}_i - \bar{y})^2$	$\frac{\sum (\hat{y}_i - \bar{y})^2}{p}$	$\frac{\frac{\sum (\hat{y}_i - \bar{y})^2}{p}}{\frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}}$
Residual	$n - p - 1$	$\sum (y_i - \hat{y}_i)^2$	$\frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}$	
Total	$n - 1$	$\sum (y_i - \bar{y})^2$		

Is At Least One Predictor Useful?

- We can use the F–statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \approx F_{p,n-p-1}$$

Quantity	Value
R^2	0.3385
F–statistic	5.756

Goodness of Fit

- How well does the model fit the data? One measure is R^2, the so-called *coefficient of determination* or *percentage of variance explained*

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y}^2)} = 1 - \frac{\text{RSS}}{\text{Total SS (corrected for mean)}}.$$

- The range is $0 \leq R^2 \leq 1$ - values closer to 1 indicating better fits. For simple linear regression $R^2 = r^2$ where r is the correlation between x and y. An equivalent definition is

$$R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{\text{Regression Sum of Squares}}{\text{Total SS (corrected for mean)}}.$$

More about R^2

- What is a good value of R^2?
- It depends on the area of application.
  - In the biological and social sciences, variables tend to be more weakly correlated and there is a lot of noise. We'd expect lower values for R^2 in these areas – a value of 0.6 might be considered good.
  - In physics and engineering, where most data comes from closely controlled experiments, we expect to get much higher R^2's and a value of 0.6 would be considered low.
- Of course, I generalize excessively here so some experience with the particular area is necessary for you to judge your R^2's well.