Multiple linear regression

Multiple linear regression

Let us recall the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

where X_j is the jth predictor and β_j quantifies the relationship between that variable and the response.
We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.

Multiple linear regression

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

The parameters are estimated through the ordinary least squares method, OLS, by minimizing

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

Multiple linear regression: assumptions on error term

We make the following assumptions regarding error terms $(\varepsilon_1, ..., \varepsilon_N)$

1. errors have mean zero
2. errors are uncorrelated
3. errors are uncorrelated with $X_{j,i}$

Multiple linear regression: model fit

The R^2 statistic is given by

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}}$$

In addition to looking at the R^2, it can be useful to plot the data. Graphical summaries may reveal problems with a model that are not visible from numerical statistics.

Multiple linear regression

In order to test the global significance of the model we

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H_1 &: \text{at least one } \beta_j \neq 0 \end{aligned}$$

through the F statistic

$$F = \frac{\text{ESS}/p}{\text{RSS}/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

Multiple linear regression

Results may be usefully displayed in an ANOVA table

Source	df	SS	MS	F
Model	p	ESS	MSR	MSR/MSE
Error	n-p-1	RSS	MSE	
Total	n-1	SST		

Multiple linear regression

After examining the global significance of the model, it is useful to evaluate the significance of parameters. The hypothesis system is

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

and the test is defined as

$$t = \frac{b_j}{\text{se}(b_j)}$$

where b_j is the estimate of the j_{th} coefficient and $\text{se}(b_j)$ is the standard error.

Multiple linear regression: collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another.

Effects of collinearity

- reduces the accuracy of estimates of the regression coefficients
- the standard error for β_j grows
- the t-statistic declines $\rightarrow$ we may fail to reject $H_0 : \beta_j = 0$

Multiple linear regression: collinearity

how do we detect a problem of collinearity?

- a simple way to detect collinearity is to look at the correlation matrix of the predictors.
- an element of this matrix that is large in absolute value indicates a pair of highly correlated variables $\rightarrow$ collinearity
- it is possible for collinearity to exist between three or more variables $\rightarrow$ multicollinearity

Multiple linear regression: collinearity

A better way to assess the multicollinearity is to compute the variance inflation factor, VIF.

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

where R_j^2 is the determination index of the regression of the j_{th} variable on the other $k-1$ predictors.

- If $R_j^2 = 0$, then $\text{VIF}_j = 1$.
- If there is a multicollinearity problem, then $\text{VIF}_j > 1$. For example, $R_j^2 = 0.9$, $\text{VIF}_j = 10$.

Example

Let us consider a sample of 10 households and the following variables:

- Y: yearly amount spent in food (hundreds eur)
- X_1: family income (thousands eur)
- X_2: number of family members

We first calculate the correlation matrix ...

	Y	X_1	X_2
Y	1	0.884	0.737
X_1		1	0.867
X_2			1

Example

We estimate the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

coefficient	estimate	std. error	t-statistic
β_0	3.51865	3.16055	1.1133
β_1	2.27762	0.81261	2.80284
β_2	-0.411406	1.23603	-0.332844

Source	df	SS	MS	F
Model	2	213.422	106.711	12.75
Error	7	58.578	8.3682	
Total	9	272		

$R^2 = 0.7846$

How do we interpret these results?

Example

Let us compute the Variance Inflation Factor.
This may be easily computed for X_1 e X_2 considering that $R^2 = (r_{X_1 X_2})^2 = (0.867)^2 = 0.75$ so that

$$\text{VIF}_{X_1} = 1/(1 - 0.75) = 4$$
$$\text{VIF}_{X_2} = 1/(1 - 0.75) = 4$$

There is a multicollinearity problem: solution $\rightarrow$ remove X_2 from the model and estimate a simple regression with X_1.