APPLIED STATISTICS

Multiple Linear Regression for Categorical Explanatory Variables

Dr Tao Zou

Research School of Finance, Actuarial Studies & Statistics
The Australian National University

Last Updated: Mon Aug 21 14:13:52 2017

Overview

- Continuous and Categorical Data
- Indicator Variables
- Interaction

References

1. **F.L. Ramsey and D.W. Schafer** (2012)
   Chapter 9 of *The Statistical Sleuth*
2. The slides are made by **R Markdown**.
   http://rmarkdown.rstudio.com

Continuous and Categorical Data

Continuous data: Corn Yield, Rainfall (in Example: Corn Yield vs Rainfall); Interval, Duration (in Example: Old Faithful).

Categorical data: Gender (male/female); Location (AUS, US, UK, NZ).

	Continuous X	Categorical X
Continuous Y	MLR	**MLR+ Indicator Variables**
Categorical Y	Logistic Regression	Logistic Regression $+$ Indicator Variables

Indicator Variables

An indicator variable takes on one of the two values: “1” indicates that an attribute is present, and “0” indicates that the attribute is absent.

For example, an indicator for gender might be set to “1” for female and “0” for male.

Indicator variables are extremely useful in regression to deal with the categorical explanatory variables.

Indicator Variable for Two-Category Explanatory Variable

Example: In a study of primary school reading ability, an educator wished to relate a measure of a student's reading ability (Y) to age (X_1) and gender (female/male).

Gender is a categorical variable with two categories.

To allow for it in MLR, we can consider the following two possible indicator variables.

$$X_2 = 1 \text{ if female; } 0 \text{ otherwise;}$$

$$X_3 = 1 \text{ if male; } 0 \text{ otherwise.}$$

It is worth noting that $X_2 + X_3 = 1$.

Indicator Variable for Two-Category Explanatory Variable (Con'd)

The following three models might then be fit:

$$\text{model } \alpha\text{: } \mu_\alpha\{Y|X\} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_3,$$

$$\text{model } \beta\text{: } \mu_\beta\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \text{ and}$$

$$\text{model } \gamma\text{: } \mu_\gamma\{Y|X\} = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3.$$

Since $X_2 + X_3 = 1$, we have

$$\begin{aligned} \mu_\alpha\{Y|X\} &= \alpha_0 + \alpha_1 X_1 + \alpha_2(1 - X_2) \\ &= (\alpha_0 + \alpha_2) + \alpha_1 X_1 + (-\alpha_2) X_2 \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \mu_\beta\{Y|X\}. \end{aligned}$$

Knowing the values of $\beta_0, \beta_1, \beta_2 \Leftrightarrow$ knowing the values of $\alpha_0, \alpha_1, \alpha_2$.

Hence model α and model β are equivalent and will give the same result.

Indicator Variable for Two-Category Explanatory Variable (Con'd)

However, we have the following for model γ,

$$\begin{aligned}\mu_\gamma\{Y|X\} &= \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 \\ &= \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3(1 - X_2) \\ &= (\gamma_0 + \gamma_3) + \gamma_1 X_1 + (\gamma_2 - \gamma_3) X_2 \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \mu_\beta\{Y|X\}.\end{aligned}$$

Hence model γ and model β are also equivalent.

But knowing the values of $\beta_0, \beta_1, \beta_2 \not\Rightarrow$ knowing the values of $\gamma_0, \gamma_1, \gamma_2, \gamma_3$.

The reason is that actually we only have the information of X_1 and X_2. The information of X_3 is redundant since $X_3 = 1 - X_2$, which is included in X_2. This phenomenon is called **multicollinearity**.

Indicator Variable for Two-Category Explanatory Variable (Con'd)

One can try that model γ cannot be fit in R.

As a consequence, drop one of the indicator variables (X_2 or X_3) and fit model α or model β.

A categorical variable with 2 categories will be represented by only $2 - 1 = 1$ indicator variable as the explanatory variable in MLR.

Indicator Variable for Two-Category Explanatory Variable – Interpretation

Since model α and model β are equivalent, we can adopt either of them. Let us consider model β.

Gender	X_2	model β	Remark
female	1	$\mu\{Y\|X\} = \beta_0 + \beta_1 X_1 + \beta_2$	
male	0	$\mu\{Y\|X\} = \beta_0 + \beta_1 X_1$	Baseline level.

Model β states that the mean reading ability is a straight line function of age for both females and males.

Marginal effect of X_1 under the category "female" is

$$\mu\{Y|X_1 = x_1 + 1, X_2 = 1\} - \mu\{Y|X_1 = x_1, X_2 = 1\} = \beta_1.$$

Marginal effect of X_1 under the category "male" is

$$\mu\{Y|X_1 = x_1 + 1, X_2 = 0\} - \mu\{Y|X_1 = x_1, X_2 = 0\} = \beta_1.$$

The above are equal. Hence, the slope of both lines is β_1.

Indicator Variable for Two-Category Explanatory Variable – Interpretation (Con'd)

However, the intercept is β_0 for males (baseline level) and $\beta_0 + \beta_2$ for females.

Marginal effect of gender changing from "male" (baseline level) to "female" is

$$\mu\{Y|X_1 = x_1, X_2 = 1\} - \mu\{Y|X_1 = x_1, X_2 = 0\} = \beta_2.$$

The coefficient β_2 is the intercept increase from the mean of response for males (baseline level) to the mean of response for females.

Symmetrically, please try to interpret the coefficients in model α by yourselves.

Indicator Variables for Multicategory Explanatory Variable

In the previous example, we now wish to relate a measure of a student's reading ability (Y) to age (X_1) and the location of the student (AUS, US, UK, NZ).

Location is a categorical variable with four categories.

To allow for it in MLR, we can consider the following four possible indicator variables.

$$X_2 = 1 \text{ if AUS; } 0 \text{ otherwise;}$$

$$X_3 = 1 \text{ if US; } 0 \text{ otherwise;}$$

$$X_4 = 1 \text{ if UK; } 0 \text{ otherwise;}$$

$$X_5 = 1 \text{ if NZ; } 0 \text{ otherwise.}$$

It is worth noting that $X_2 + X_3 + X_4 + X_5 = 1$.

Indicator Variables for Multicategory Explanatory Variable (Con'd)

Similarly, we only have the information of X_1, X_2, X_3 and X_4. The information of X_5 is redundant since $X_5 = 1 - X_2 - X_3 - X_4$, which is included in X_2, X_3 and X_4.

As a consequence, drop X_5 and fit MLR model

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

A categorical variable with C categories will be represented by only $C - 1$ indicator variables as the explanatory variables in MLR.

Indicator Variables for Multicategory Explanatory Variable – Interpretation

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Location	X_2	X_3	X_4	model	Remark
AUS	1	0	0	$\mu\{Y\|X\} = (\beta_0 + \beta_2) + \beta_1 X_1$	
US	0	1	0	$\mu\{Y\|X\} = (\beta_0 + \beta_3) + \beta_1 X_1$	
UK	0	0	1	$\mu\{Y\|X\} = (\beta_0 + \beta_4) + \beta_1 X_1$	
NZ	0	0	0	$\mu\{Y\|X\} = \beta_0 + \beta_1 X_1$	Baseline level.

Marginal effect of X_1 under all categories is β_1. Hence, the slope of all lines is β_1.

However, the intercept is β_0 for NZ (baseline level), $\beta_0 + \beta_2$ for AUS, $\beta_0 + \beta_3$ for US, and $\beta_0 + \beta_4$ for UK, respectively.

Note that the baseline level corresponds to

$$X_5 = 1 \text{ if NZ; } 0 \text{ otherwise,}$$

which is **not** used in the above MLR model.

Indicator Variables for Multicategory Explanatory Variable – Interpretation (Con'd)

Marginal effect of location changing from "AUS" to "NZ" (baseline level) is $\mu\{Y|X_1 = x_1, X_2 = 1, X_3 = 0, X_4 = 0\} - \mu\{Y|X_1 = x_1, X_2 = 0, X_3 = 0, X_4 = 0\} = \beta_2$.

The coefficient β_2 is the intercept increase from the mean of response for "NZ" (baseline level) to the mean of response for "AUS".

Marginal effect of location changing from "US" to "NZ" (baseline level) is $\mu\{Y|X_1 = x_1, X_2 = 0, X_3 = 1, X_4 = 0\} - \mu\{Y|X_1 = x_1, X_2 = 0, X_3 = 0, X_4 = 0\} = \beta_3$.

Marginal effect of location changing from "UK" to "NZ" (baseline level) is $\mu\{Y|X_1 = x_1, X_2 = 0, X_3 = 0, X_4 = 1\} - \mu\{Y|X_1 = x_1, X_2 = 0, X_3 = 0, X_4 = 0\} = \beta_4$.

Product Terms for Interaction

We now go back to the original example and wish to relate a measure of a student's reading ability (Y) to age (X_1) and gender (female/male).

Since a categorical variable with C categories will be represented by only $C-1$ indicator variables as the explanatory variables in MLR, we only use

$$X_2 = 1 \text{ if female; } 0 \text{ otherwise;}$$

in MLR model.

However, we consider the model with a product term

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2.$$

Product Terms for Interaction – Interpretation

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2.$$

Gender	X_2	model	Remark
female	1	$\mu\{Y|X\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1$	
male	0	$\mu\{Y|X\} = \beta_0 + \beta_1 X_1$	Baseline level.

Marginal effect of X_1 under the category “female” is

$$\mu\{Y|X_1 = x_1 + 1, X_2 = 1\} - \mu\{Y|X_1 = x_1, X_2 = 1\} = \beta_1 + \beta_3.$$

Marginal effect of X_1 under the category “male” (baseline level) is

$$\mu\{Y|X_1 = x_1 + 1, X_2 = 0\} - \mu\{Y|X_1 = x_1, X_2 = 0\} = \beta_1.$$

The above are not equal. Hence, the slopes of two lines are different.

“Age” and “gender” can be said to **interact** if the (marginal) effect of “age” that has on the mean of response depends on the value of “gender”.

The product term $X_1 \times X_2$ models an **interaction** between age and gender.

Product Terms for Interaction – Interpretation (Con'd)

The coefficient β_3 is the slope increase from the mean of response for males (baseline level) to the mean of response for females.

Marginal effect of gender changing from "male" (baseline level) to "female" is

$$\mu\{Y|X_1 = x_1, X_2 = 1\} - \mu\{Y|X_1 = x_1, X_2 = 0\} = \beta_2 + \beta_3 x_1.$$

Here the (marginal) effect of "gender" that has on the mean of response depends on the value of "age" (x_1).

The coefficient β_2 is the intercept increase from the mean of response for males (baseline level) to the mean of response for females.

Product Terms for Interaction – Interpretation (Con'd)

Now, both the intercept and slope depend on gender.

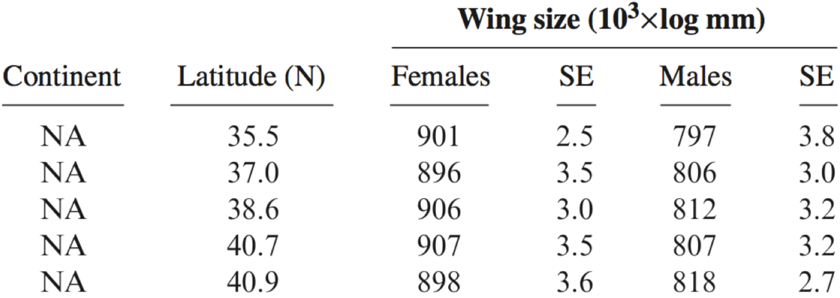
Often it is very necessary to test whether the interaction is present ($\beta_3 \neq 0$).

Note: Except in special circumstances, a model including a product term for interaction between two explanatory variables should also include terms with each of the explanatory variables individually.

Example: Size of Wing

(example from “The Statistical Sleuth”)

This dataset contains information on a type of fly (Drosphila subobscura). Researchers were interested in whether wing size was related to location and/or latitude.



		Wing size (10^3×log mm)			
Continent	Latitude (N)	Females	SE	Males	SE
NA	35.5	901	2.5	797	3.8
NA	37.0	896	3.5	806	3.0
NA	38.6	906	3.0	812	3.2
NA	40.7	907	3.5	807	3.2
NA	40.9	898	3.6	818	2.7

Example: R Code

```r
rm(list=ls())
setwd('~/Desktop/Research/AppliedStat2017/L5')
#install.packages('Sleuth3')
library(Sleuth3)
wing=ex0918
wingsize=wing$Females   #response variable
con=wing$Continent      #explan variable (categorical)
lat=wing$Latitude       #explan variable quantitative
con
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA NA EU EU EU EU EU EU EU EU EU EU
## Levels: EU NA
```

```r
#creating the indicator for North America
indNA=ifelse(con=="NA",1,0)
indNA
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
```

```r
#fitting the MLR allowing for an interaction
wingint.reg=lm(wingsize~lat+indNA+indNA*lat)
```

$$\mu\{Y|X\} = \beta_0 + \beta_1 \text{lat} + \beta_2 \text{indNA} + \beta_3 \text{lat} \times \text{indNA}.$$

Continent	indNA	model	Remark
NA	1	$\mu\{Y\|X\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{lat}$	
EU	0	$\mu\{Y\|X\} = \beta_0 + \beta_1 \text{lat}$	Baseline level.

R Code (Con'd)

```r
summary(wingint.reg)
```

```
##
## Call:
## lm(formula = wingsize ~ lat + indNA + indNA * lat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.355  -2.332   0.384   5.434  21.294
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 836.1904    28.3200  29.527 4.76e-16 ***
## lat           1.7838     0.6105   2.922  0.00951 **
## indNA       -16.8939    40.5453  -0.417  0.68214
## lat:indNA     0.3109     0.9032   0.344  0.73486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.13 on 17 degrees of freedom
## Multiple R-squared:  0.5568, Adjusted R-squared:  0.4785
## F-statistic: 7.118 on 3 and 17 DF,  p-value: 0.002643
```

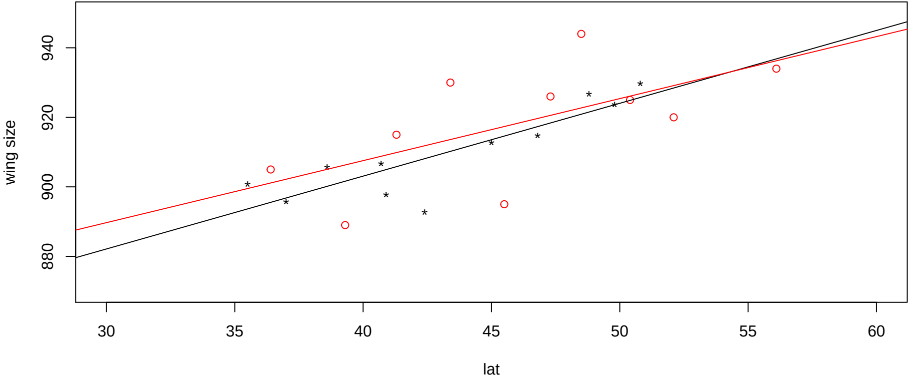
Interpretation:

1. If Continent changes from “EU” (baseline level) to “NA”, the estimated slope of the mean of response will increase 0.3109 unit.

2. If Continent changes from “EU” (baseline level) to “NA”, the estimated intercept of the mean of response will decrease 16.8939 unit.

R Code (Con'd)

```
#plotting points from NA
plot(lat[indNA==1],wingsize[indNA==1],pch="*",ylab="wing size", xlab="lat",xlim=c(30,60),ylim=c(870,950))
#adding the points from EU (Baseline Level)
points(lat[indNA==0],wingsize[indNA==0],col='red')
#regression line for NA
abline(wingint.reg$coef[1]+wingint.reg$coef[3],wingint.reg$coef[2]+wingint.reg$coef[4])
#regression line for EU (Baseline Level)
abline(wingint.reg$coef[1],wingint.reg$coef[2],col='red')
```



Graphically, the interaction does not appear to be important.

Comments on R Output

The regression coefficients of both "indNA" and "lat×indNA" are not significant at the significance level 0.05.

1. Can we eliminate both variables in the model immediately? NO! We will explain the reason in the inferential tools for multiple linear regression.

2. Can we eliminate one of them? Which one should be eliminated? We will answer this quesiton in the variable selection lectures.

3. Based on the "summary()" output, the LS squares estimates are $\hat{\beta}_0 = 836.19$, $\hat{\beta}_1 = 1.78$, $\hat{\beta}_2 = -16.89$ and $\hat{\beta}_3 = 0.31$. Since both β_2 and β_3 are not significantly different from 0, can we just let $\hat{\beta}_2 = 0$ and $\hat{\beta}_3 = 0$? NO!

Comments on R Output (Con'd)

There are two reasons:

(1). $\hat{\beta}_0 = 836.19$, $\hat{\beta}_1 = 1.78$, $\hat{\beta}_2 = -16.89$ and $\hat{\beta}_3 = 0.31$ are LS squares estimates such that $Q(b_0, b_1, b_2, b_3)$ in Lecture Notes 4 is minimised, which corresponds to "best fitting". However, if we let $\hat{\beta}_2 = 0$ and $\hat{\beta}_3 = 0$, those no longer minimise $Q(b_0, b_1, b_2, b_3)$ and do not achieve "best fitting".

(2). Even if the regression coefficients of both indNA and lat×indNA are not significant, we cannot accept $\beta_2 = 0$ and $\beta_3 = 0$ based on Lecture Notes 2. That means $\beta_2 \neq 0$ and $\beta_3 \neq 0$ are possible but we just do not have enough evidence to show it.

4. If we determine to eliminate one explanatory variables in the MLR, we need to refit the model with the explanatory variables after elimination, and obtain all new LS estimates for the regression coefficients.