1 Multiple Linear Regression

In *multiple linear regression*, we use the values of more than one explanatory variable to predict or describe the values of a response variable. That is, the equation for the mean of the response variable (y) is a function of two or more explanatory variables.

Ex: Suppose a researcher wanted to predict the GPA of students, but wanted to account for several variables such as the following

x_1	=	Study hours	=	Hours spent studying per week
x_2	=	Classes missed	=	Number of classes student misses in typical week
x_3	=	Work hours	=	Hours per week that student works in a part-time or full-time job

The general form of a *linear multiple regression model* relating grade point average (GPA) to these three predictor variables is

$$\text{GPA} = \beta_0 + \beta_1 \text{Study hours} + \beta_2 \text{Classes missed} + \beta_3 \text{Work hours}$$

Numerical estimates of the parameters β_0, β_1, β_2, and β_3 would be determined by using data from a sample of students for whom information on grade point average and the three predictor variables is available.

In general, suppose y is the response variable, and $x_1, x_2, \ldots, x_p$ is a set of explanatory variables. The *linear multiple regression equation* for the relationship between the mean value of y and the variables $x_1, x_2, \ldots, x_p$ is

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

- $E(Y)$ represents the mean value of y for individuals in the population who all have the same particular values of $x_1, x_2, \ldots, x_p$.
- $\beta_0, \beta_1, \ldots, \beta_p$ are referred to as the *regression coefficients* or, alternatively, may be called either the β *coefficients* or the β *parameters*.

Finding the β_i's "by hand" is beyond the scope of this class. It involves matrix algebra. We can, however, use Excel to find their values.

Ex: The temperature dataset gives geographic latitude, mean January temperature, mean April temperature, and mean August temperature for 20 cities in the United States. The temperatures are recorded as degrees Fahrenheit. A multiple linear regression equation for the relationship between August temperature and the explanatory variables latitude, January temperature, and April temperature is

$$E\,(\text{August temp}) = \beta_0 + \beta_1 \text{Latitude} + \beta_2 \text{January temp} + \beta_3 \text{April temp}$$

To find the β_i's, click the Data tab, and then Data Analysis. Choose Regression, click OK. We choose the data values in the column for "Aug Temp" for our Input Y Range. We choose data values in the three columns "latitude", "Jan Temp", and "Apr Temp" for our Input X Range. Click OK.

Excel puts the output on Sheet 2. Look for *Coefficients* in the last row of data. The "Intercept" is the $\beta_0 \approx -29$. You see "X Variable" repeated with numbers in the adjacent column. By the ways that our columns were arranged, this tells us that $\beta_1 \approx 0.70$, $\beta_2 \approx -0.44$, ond $\beta_3 \approx 1.66$. So our sample regression equation is

$$y = -29 + 0.70x_1 - 0.44x_2 + 1.66x_3$$

Note here that

$$y = \text{August temp} \quad x_1 = \text{Latitude} \quad x_2 = \text{January temp} \quad x_3 = \text{April temp}$$

Some correlations are

	Latitude	January temp	April temp
August temp	-0.78072879	0.62184167	0.84615341
Latitude	1	-0.8559871	-0.9576257
January temp	-0.8559871	1	0.91125924

Here are some oddities:

By looking at the coefficients in the linear equation, you see a positive coefficient for x_1, which would indicate that the August temperature goes up as latitude increases. This does not seem reasonable. Regardless of month, average temperature tends to decrease as latitude increases in the Northern hemisphere. We see that with the correlation coefficient.

The value of the coefficient for x_2 is negative, indicating that the warmer it is in January, the colder it is in August. This does not seem reasonable.

Why is this so?

With such strong associations among the explanatory variables (see chart above), it becomes nearly impossible to separate their individual effects on the response variable. Each coefficient actually represents the additional contribution made by that variable, given that all of the others are already being used as predictors. The moral of the story is that caution must be employed in interpreting the individual regression coefficients in a multiple regression analysis.

Ex: Let us predict Phoenix using the data in the dataset. (We use $\hat{y}$ to denote the predicted value.) We have

$$\hat{y} = -29 + 0.70\,(33) - 0.44\,(54) + 1.66\,(70) = 86.5 \text{ degrees}$$

but the actual recorded degree was $y = 92$. (We use y for actual value.) Our error is $e = y - \hat{y} = \text{actual} - \text{predicted} = 92 - 86.5 = 5.5$ degrees.

The statistic denoted by R^2 measures what researchers often term the "proportion of variation in the response explained by the predictor variables (the x variables)."

In layman's terms, if R^2 is closer to 1, then using $x_1, x_2, \ldots, x_p$ to predict y is reasonable. If R^2 is closer to zero, then you are better off just using the mean of y and SD_y to predict y.

When working multiple linear regression, you should use the Adjusted R Square from your Excel output. It is more trusted than the R Square that is listed in the output.

2 Hypothesis Testing and Multiple Linear Regression

The hypotheses are

$$H_0 : \beta_k = 0 \qquad \text{and} \qquad H_a : \beta_k \neq 0$$

for each β_k in the model. Notice that if $\beta_k = 0$ for a β coefficient that multiplies an explanatory variable, then the variable x_k is not influencing the calculation of predicted y-values because $\beta_k x_k = 0$. Thus, these tests are used to make decisions about which explanatory variables may or may not be helpful predictors of the response variable.

Ex (cont.): For our August example, we look at our Excel output see that there are p-values for (using a level of significance of 0.5 in decimal instead of 5%)

Latitude	$\beta_1 = 0$	p-value $= 0.102987$	not statistically significant
January temp	$\beta_2 = 0$	p-value $= 0.00062$	statistically significant
April temp	$\beta_3 = 0$	p-value $= 0$	statistically significant

There is a strong connection between latitude and August temperature, but January and April temperatures, which themselves are related to latitude, provide sufficient information to predict August temperature. The consequence is that we could consider removing the latitude variable from the model because it is not making a significant additional contribution to the prediction of y. If we do so, however, we must again use statistical software to estimate the regression coefficients for the model with only two predictor variables.

The test for $\beta_0 = 0$ (information given in the row labeled Constant) is not useful in this situation. The coefficient β_0 corresponds to the August temperature when the value of each explanatory variable is 0. That is not a realistic combination of x-values.

Adapted from supplemental chapters from *Mind on Statistics, 3rd Ed* by Utts Heckard.

Mail to rstephens@colgate.edu