6.1 Multiple Linear Regression

Wednesday, 26 October 2022 13:07

Summary	• Multiple Linear Regression • $\bar{R}^2$ and s_e • Calibration Plot • Marginal and Partial Slopes • Path Diagram and Collinearity • F-statistic • t-statistic • Prediction interval
• What's the basic difference between multiple and simple linear regression?	**Multiple Regression** • In simple linear regression, there's one explanatory variable and one response (dependent) variable. • In multiple linear regression, we have multiple explanatory variables and one response variable.
	The Multiple Regression Model ▪ Use multiple regression to describe the relationship between several explanatory variables and the response. ▪ Multiple regression separates the effects of each explanatory variable on the response and reveals which really matter. • We will study the effect of each of these explanatory variable on the response variable.

• Full form of MLR and MRM.
• What is k in MRM?
	The Multiple Regression Model
	▪ Multiple regression model (MRM): model for the association in the population between multiple explanatory variables and a response.
	▪ k: the number of explanatory variables in the multiple regression (k = 1 in simple regression).
	• We can call it either: MLR / MRM.
	  ○ MLR : Multiple Linear Regression
	  ○ MRM: Multiple Regression Model

• Equation of MRM • Assumptions about the error term	**The Multiple Regression Model** The response Y is linearly related to k explanatory variables X_1, X_2, and X_k by the equation $Y = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \epsilon$ $\epsilon \sim N(0, \sigma_\epsilon^2)$ ▪ *The unobserved errors in the model* 1. *are independent of one another,* 2. *have equal variance, and* 3. *are normally distributed around the regression equation.* • We will estimate β_0, β_1, $---$,β_k .
• Difference in the error terms between MLR and SLR	**The Multiple Regression Model** ▪ While the SRM bundles all but one explanatory variable into the error term, multiple regression allows for the inclusion of several variables in the

	model. ▪ In the MRM, residuals departing from normality may suggest that an important explanatory variable has been omitted. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \quad - - - \quad + \beta_k X_k + \varepsilon$ • The impact of all other explanatory variables, which was not considered in SLR, will get lumped into the error term of SLR.
• $\bar{R}^2$ • s_e • How do they behave when an explanatory variable is added? ☐? If $\bar{R}^2 < R^2$, and $\bar{R}^2$ keeps increasing as you add more explanatory variables, then does this $\bar{R}^2 < R^2$ relationship always hold? Does it mean that no matter how many explanatory variables you add, you'll never cross R^2 ?	# Interpreting Multiple Regression ▪ R-squared and s_e ▪ $\bar{R}^2$ is known as the adjusted *R*-squared. It adjusts for both sample size *n* and model size *k*. It is always smaller than R^2. ▪ The residual degrees of freedom (*n-k-1*) is the divisor of s_e. $\bar{R}^2$ and s_e move in opposite directions when an explanatory variable is added to the model ($\bar{R}^2$ goes up while s_e goes down). $Adjusted\ R^2 < R^2$ • Adjusted R^2 gives us a slightly more realistic picture of what is the combined explanatory power of all these explanatory variables. s_e is the esimate of [Equation] Standard Deviation of the error term, standard error in estimating the slopes • In general, we want a large value of [Equation] , and we want a smaller value of s_e ,*i.e.*, if you add explanatory variables to the model: ○ [Equation] and [Equation]
• State Adjusted R squared formula	Adjusted R squared formula (not in the lectures but is asked in the questions) [Equation]

• Define Calibration Plot
• What does [Equation] represent in:
  ○ SLR
  ○ MLR
	# Interpreting Multiple Regression
	
	▪ Calibration Plot
	
	▪ Calibration plot: scatterplot of the response y on the fitted values $\hat{y}$.
	
	▪ R^2 is the correlation between $\hat{y}$ and y; the tighter data cluster along the diagonal line in the calibration plot, the larger the R^2 value.
	
	• In SLR, R represented the correlation between [Equation] and Y.
	
	• In MLR, R represents the coefficent of correlation between observed value(y) of the response variable and the fitted/predicted value([Equation]) of the response variable.

• Define
  ○ Marginal slope
  ○ Partial slope
• When is Marginal slope [Equation] Partial slope?
	# Interpreting Multiple Regression
	
	▪ Marginal and Partial Slopes
	
	▪ Partial slope: slope of an explanatory variable in a multiple regression that statistically excludes the effects of other explanatory variables.
	
	▪ Marginal slope: slope of an explanatory variable in a simple regression.

- Partial and marginal slopes only agree when the explanatory variables are uncorrelated.

- In SLR: [Equation] Marginal Slope.
  - Change in Y variable with one unit change in X variable.
- In MLR: [Equation] Partial Slope.
  - Change in Y variable with one unit change in X variable keeping all the other X variables constant.
- Ideally, we want explanatory variables which are orthogonal to each other, *i.e.*, which are independent of each other.
  - If the variables are truly independent of each other, then the marginal slope and the partial slope will have the same value. But it happens rarely.

- What is a path diagram?
  - What is the total effect of X on Y?
  - It is represented by the _____ slope?
- Define collinearity

Interpreting Multiple Regression

Path Diagram

- Path diagram: schematic drawing of the relationships among the explanatory variables and the response.

- Collinearity: very high correlations among the explanatory variables that make the estimates in multiple regression uninterpretable.

- Collinearity: When the explanatory variables are not independent, *i.e.*, when they are correlated. It is a situation that represents very high correlation amongst the explanatory variables. Sometimes it is so severe that it actually makes the estimates of MLR very difficult to interpret.
- Path Diagram:
- 

- We say that [Equation] impacts Y in two different ways:
  - There's a direct effect of [Equation] on Y, and
  - There's also an indirect effect of [Equation] on [Equation] as it impacts X_2, because they are correlated, that in turn imapcts Y.
  - [Equation]
- Total effect of [Equation] on [Equation] is represented in the Marginal Slope.

Slides that were not covered in lectures

- Errors in MRM satisfy what conditions?

Checking Conditions

Conditions for Inference

- Use the residuals from the fitted MRM to check that the errors in the model

- are independent;
- have equal variance; and
- follow a normal distribution.

- How do you calculate F-Statistic in MRM?
- What null hypothesis do we assume in MRM?

Inference in Multiple Regression

Inference for the Model: *F*-test

- The *F*-Statistic

$$F = \frac{R^2}{1 - R^2} \frac{n - k - 1}{k}$$

	is used to test the null hypothesis that all slopes are equal to zero, $$H_0: \beta_1 = \beta_2 = 0$$ [Equation] [Equation]
• What is the purpose of t-statistic in MRM? • How is t-statistic calculated?	Inference in Multiple Regression Inference for One Coefficient ▪ The *t*-statistic is used to test each slope using the null hypothesis H_0: $\beta_j = 0$. ▪ The *t*-statistic is calculated as $$t_j = \frac{b_j - 0}{se(b_j)}$$ [Equation]
• An approximate [Equation] prediction interval is given by?	Inference in Multiple Regression • Prediction Intervals ▪ An approximate 95% prediction interval is given by $\hat{y} \pm 2s_e$. ▪ For example, the 95% prediction interval for price.