Multiple Regression

The Multiple Regression Model

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept

Population slopes

Random Error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

Estimated (or predicted) value of y

Estimated intercept

Estimated slope coefficients

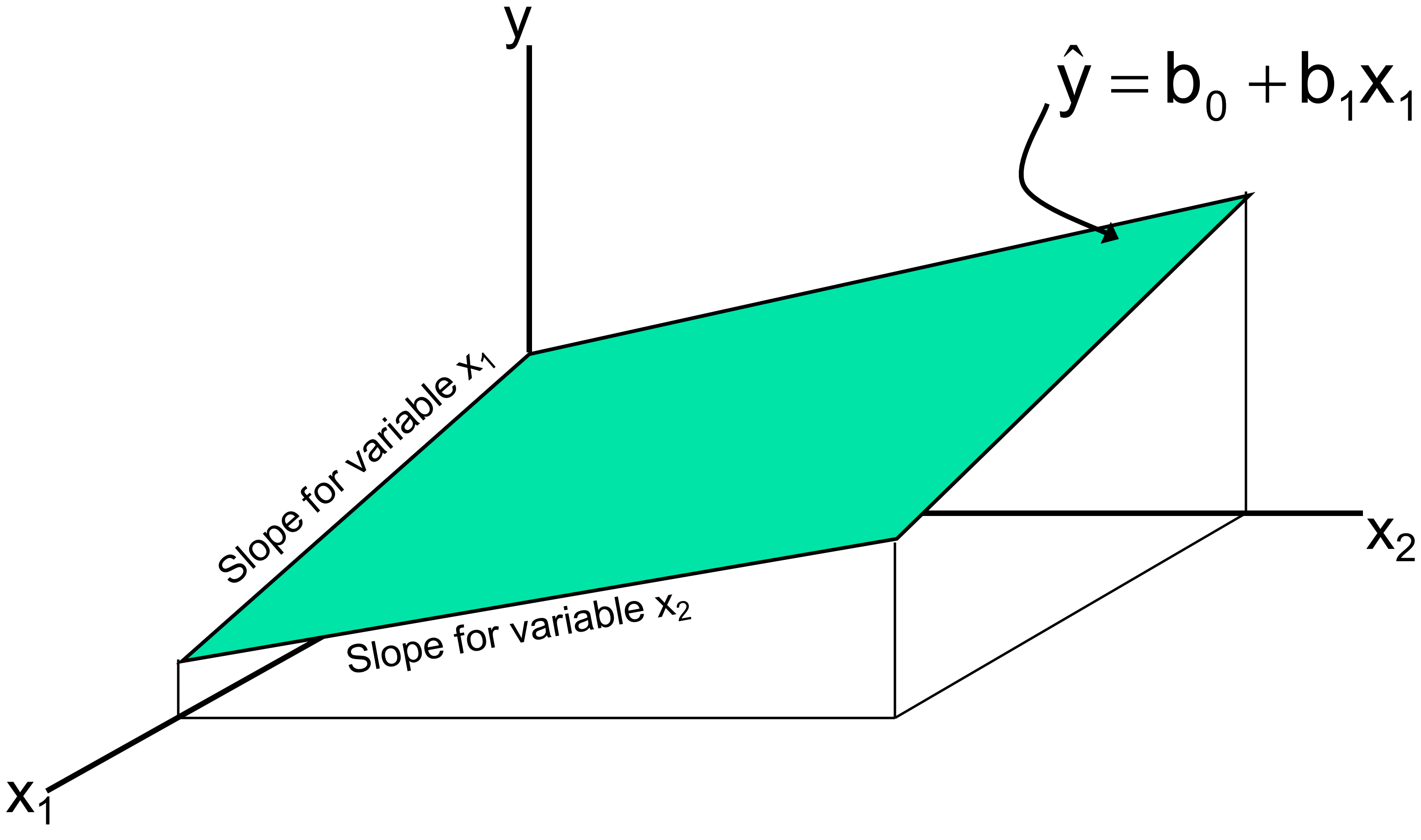
$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}$$

Multiple Regression Equation

(continued)

Two variable model

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$



Standard Multiple Regression Assumptions

- The values x_i and the error terms ε_i are independent

- The error terms are random variables with mean 0 and a constant variance, σ^2.

$$\boxed{E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \qquad \text{for } (i = 1, \ldots, n)}$$

(The constant variance property is called homoscedasticity)

Example:
2 Independent Variables

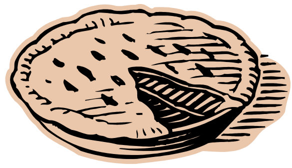
- A distributor of frozen desert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in $)
    Advertising ($100's)
- Data are collected for 15 weeks

Pie Sales Example

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1\,(\text{Price}) + b_2\,(\text{Advertising})$$



The Multiple Regression Equation

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

where
Sales is in number of pies per week
Price is in $
Advertising is in $100's.

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price

Coefficient of Determination, R^2

- Reports the proportion of total variation in y explained by all x variables taken together

$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

- This is the ratio of the explained variability to total sample variability

Prediction

- Given a population regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

- then given a new observation of a data point

  $(x_{1,n+1},\ x_{2,n+1},\ \ldots,\ x_{K,n+1})$

  the best linear unbiased forecast of $\hat{y}_{n+1}$ is

$$\hat{y}_{n+1} = b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_K x_{K,n+1}$$

- It is risky to forecast for new X values outside the range of the data used to estimate the model coefficients, because we do not have data to support that the linear model extends beyond the observed range.

Using The Equation to Make Predictions

Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

$$= 306.526 - 24.975\,(5.50) + 74.131\,(3.5)$$

$$= 428.62$$

Predicted sales is 428.62 pies

Note that Advertising is in \$100's, so \$350 means that X_2 = 3.5

Dummy Variables

- A dummy variable is a categorical independent variable with two levels:
  - yes or no, on or off, male or female
  - recorded as 0 or 1
- Regression intercepts are different if the variable is significant
- Assumes equal slopes for other variables
- If more than two levels, the number of dummy variables needed is (number of levels - 1)

Dummy Variable Example

$$\boxed{\hat{y} = b_0 + b_1 x_1 + b_2 x_2}$$

Let:

y = Pie Sales

x_1 = Price

x_2 = Holiday (X_2 = 1 if a holiday occurred during the week)
(X_2 = 0 if there was no holiday that week)

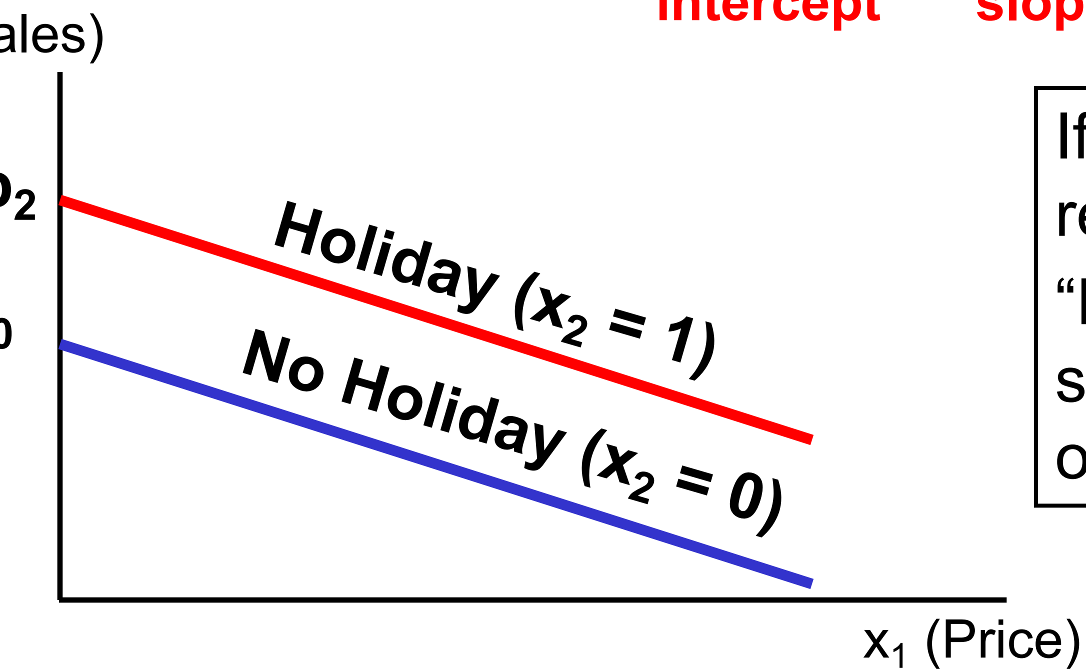
Dummy Variable Example

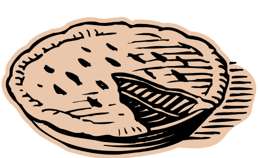
(continued)

$$\hat{y} = b_0 + b_1 x_1 + b_2(1) = \boxed{(b_0 + b_2)} + \boxed{b_1} x_1 \quad \textbf{Holiday}$$

$$\hat{y} = b_0 + b_1 x_1 + b_2(0) = \boxed{b_0} + \boxed{b_1} x_1 \quad \textbf{No Holiday}$$

Different intercept **Same slope**





If $H_0: \beta_2 = 0$ is rejected, then "Holiday" has a significant effect on pie sales

Interpreting the Dummy Variable Coefficient

Example: $\text{Sales} = 300 - 30(\text{Price}) + 15(\text{Holiday})$

Sales: number of pies sold per week

Price: pie price in \$

$$\text{Holiday:} \begin{cases} 1 & \text{If a holiday occurred during the week} \\ 0 & \text{If no holiday occurred} \end{cases}$$

$b_2 = 15$: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

Multiple Regression Assumptions

Errors (residuals) from the regression model:

$$e_i = (y_i - \hat{y}_i)$$

Assumptions:

- The errors are normally distributed
- Errors have a constant variance
- The model errors are independent

Thank You!