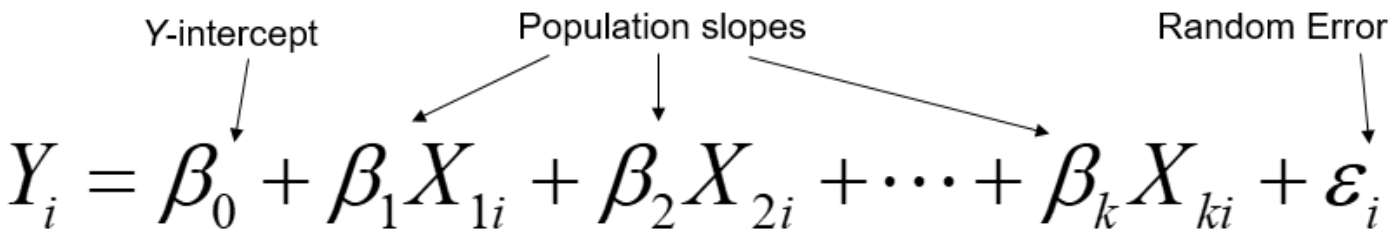
# Topic 11: Multiple Regression Analysis

Dr. Lei Pan

School of Accounting, Economics and Finance
Curtin University

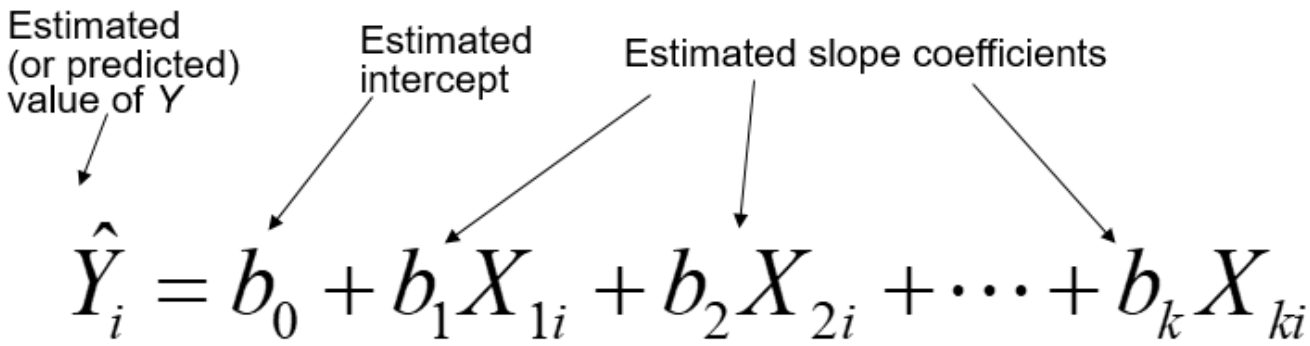
# The Multiple Regression Model

- Idea: Examine the linear relationship between one dependent variable ($Y$) and two or more independent variables ($X_i$).
- **Multiple Regression Model with k independent variables**:



# Multiple Regression Equation (1 of 2)

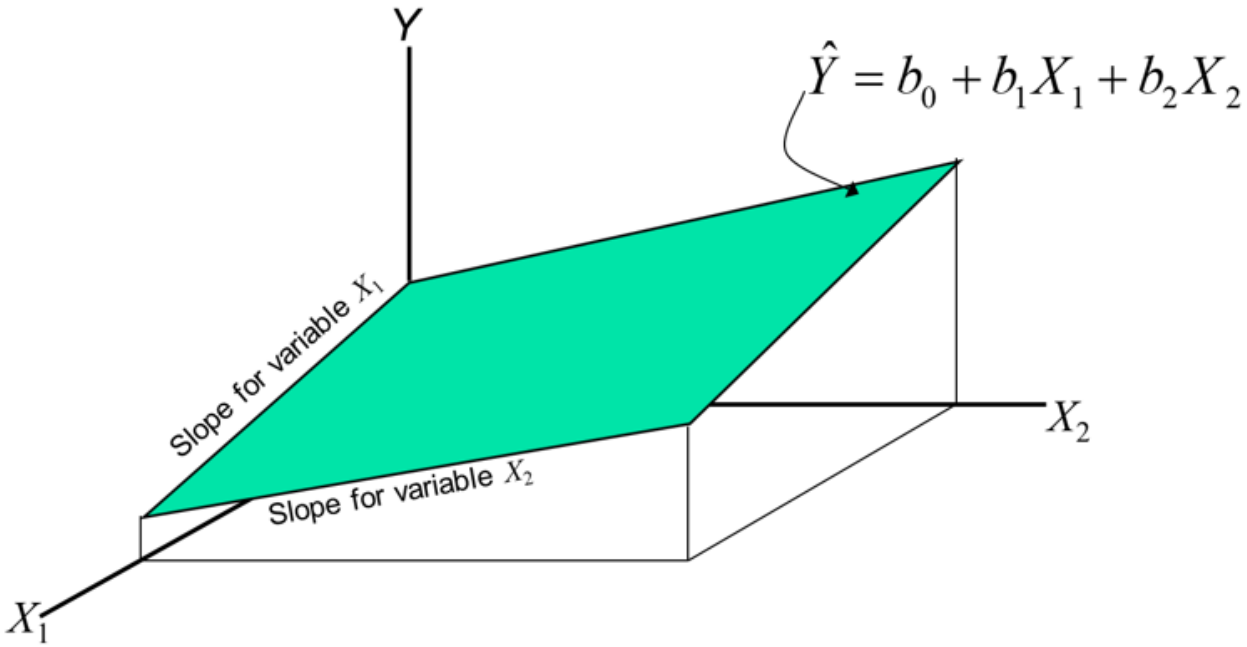
- The coefficients of the multiple regression model are estimated using sample data.
- **Multiple regression equation with k independent variables**:




- We will use Excel to obtain the regression slope coefficients and other regression summary measures.

# Multiple Regression Equation (2 of 2)

**Two variable model**



# Example: Two Independent Variables

- A distributor of frozen dessert 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:**

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

Curtin University

# Multiple Regression Excel Output

| Regression Statistics | |
|---|---|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |



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

| ANOVA | df | SS | MS | F | Significance F |
|---|---|---|---|---|---|
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 |
| Residual | 12 | 27033.306 | 2252.776 | | |
| Total | 14 | 56493.333 | | | |

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | −24.97509 | 10.83213 | −2.30565 | 0.03979 | −48.57626 | −1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

Curtin University

50 YEARS OF INNOVATION

# The Multiple Regression Equation

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

- $\mathbf{b_1 = -24.975}$: 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
- $\mathbf{b_2 = 74.131}$: 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

# Using the Equation to Make Predictions

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

$$\begin{aligned} \text{Sales} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975\,(5.50) + 74.131\,(3.5) \\ &= 428.62 \end{aligned}$$

**Predicted sales is 428.62 pies**

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

# Coefficient of Multiple Determination, $r^2$

- Reports the proportion of total variation in $Y$ explained by all $X$ variables taken together.

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

# Coefficient of Multiple Determination in Excel

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

| ANOVA | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 29460.027 | 14730.01 | 6.53861 | 0.01201 |
| Residual | 12 | 27033.306 | 2252.776 | | |
| Total | 14 | 56493.333 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

Curtin University

50 YEARS OF INNOVATION

# You Do It

A two predictor regression equation using total SAT score (SATSUM) and High-school grades (HSGPA) to predict first-year college grades (FYGPA):

$$\text{predicted FYGPA} = -0.873 + 0.00144\text{SATSUM} + 0.58\text{HSGPA}$$

$R^2 = 0.358$

What percentage of the variability in first-year college grades are NOT explained by SATSUM and HSGPA?

# Adjusted $r^2$ (1 of 2)

- $r^2$ never decreases when a new $X$ variable is added to the model
  - This can be a disadvantage when comparing models.
- What is the net effect of adding a new variable?
  - We lose a degree of freedom when a new $X$ variable is added.
  - Did the new $X$ variable add enough explanatory power to offset the loss of one degree of freedom?

# Adjusted $r^2$ (2 of 2)

- Shows the proportion of variation in $Y$ explained by all $X$ variables adjusted for the number of $X$ variables used.

$$r^2_{adj} = 1 - [(1 - r^2)(\frac{n-1}{n-k-1})]$$

(where $n$ = sample size, $k$ = number of independent variables)

  - Penalizes excessive use of unimportant independent variables
  - Smaller than $r^2$
  - Useful in comparing among models

# Adjusted $r^2$ in Excel

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

| ANOVA | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 29460.027 | 14730.01 | 6.53861 | 0.01201 |
| Residual | 12 | 27033.306 | 2252.776 | | |
| Total | 14 | 56493.333 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

Curtin University

50 YEARS OF INNOVATION

# Is the Model Significant?

- $F$ test for overall significance of the model.
- Shows if there is a linear relationship between all of the $X$ variables considered together and $Y$.
- Use $F$-test statistic.
- Hypotheses:
  - $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ (no linear relationship)
  - $H_1$ : at least one $\beta_i \neq 0$ (at least one independent variable affects $Y$)

Curtin University

# $F$ Test for Overall Significance (1 of 3)

- Test statistic:

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

where $F$ has:

  - numerator $= k$ degrees of freedom
  - denominator $= (n - k - 1)$ degrees of freedom

# $F$ Test Excel Output

| Regression Statistics | |
|---|---|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

*P*-value for the *F* Test

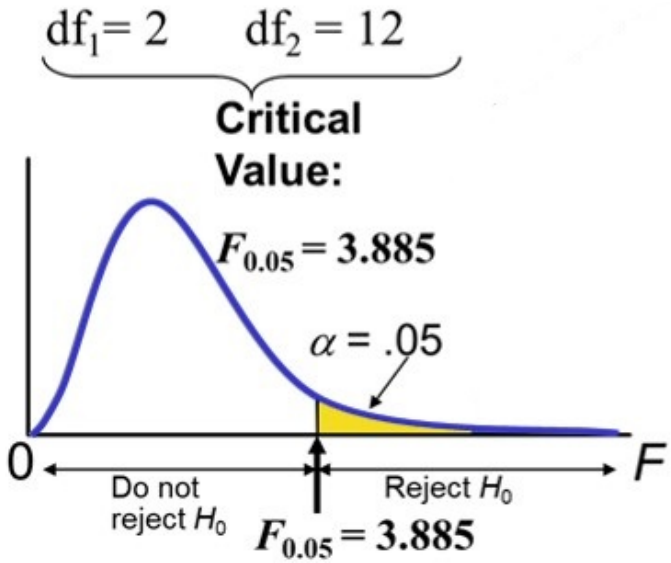
| ANOVA | df | SS | MS | F | Significance F |
|---|---|---|---|---|---|
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 |
| Residual | 12 | 27033.306 | 2252.776 | | |
| Total | 14 | 56493.333 | | | |

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

Curtin University

# $F$ Test for Overall Significance (2 of 3)

- $H_0 : \beta_1 = \beta_2 = 0$
  $H_1$: $\beta_1$ and $\beta_2$ not both zero
- $\alpha = 0.05$
- Test statistic: $F = \frac{MSR}{MSE} = 6.5386$

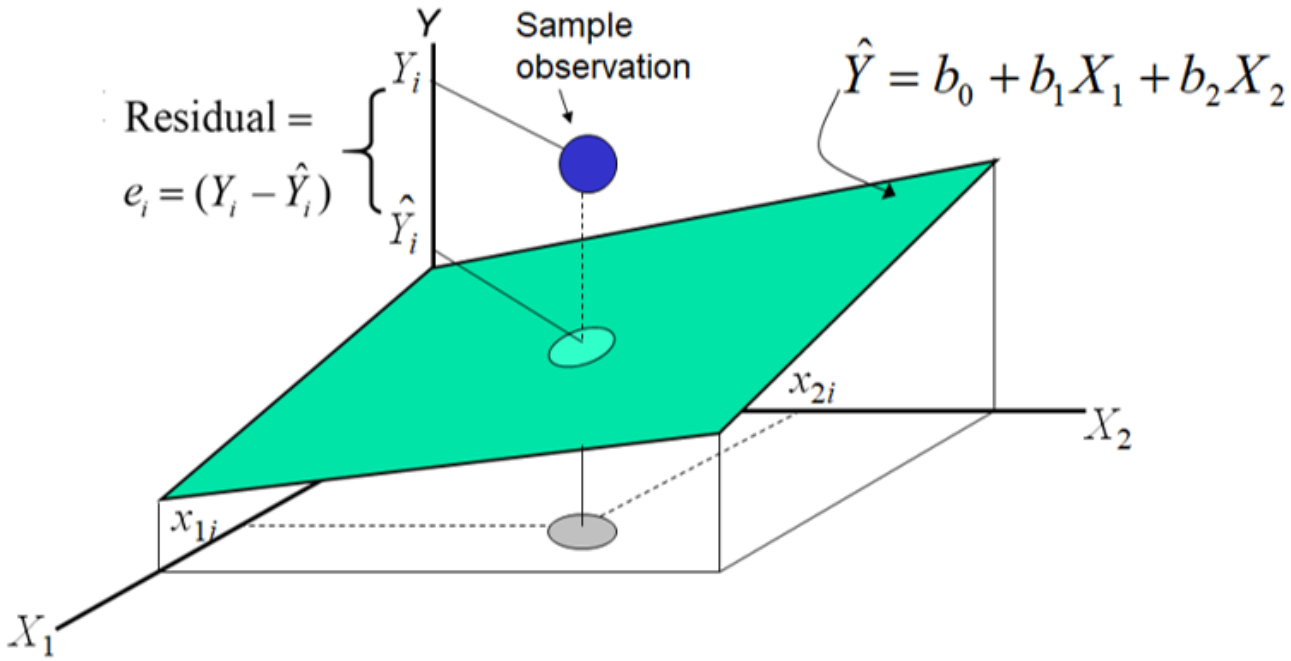


# $F$ Test for Overall Significance (3 of 3)

- Decision: Since $F$ test statistic is in the rejection region ($p$-value $< 0.05$), reject $H_0$.
- Conclusion: There is evidence that at least one independent variable affects $Y$.

# Residuals in Multiple Regression

- The best fit equation, $\hat{Y}$, is found by minimising the sum of squared errors, $\sum e^2$.



# 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.

# Residual Plots Used in Multiple Regression

- These residual plots are used in multiple regression:
  - Residuals vs. $\hat{Y}_i$
  - Residuals vs. $X_{1i}$
  - Residuals vs. $X_{2i}$
  - Residuals vs. time (if time series data)
  - Use the residual plots to check for violations of regression assumptions

# Are Individual Variables Significant?

- Use $t$ tests of individual variable slopes.
- Shows if there is a linear relationship between the variable $X_j$ and $Y$ holding constant the effects of other $X$ variables.
- Hypotheses:
  - $H_0 : \beta_j = 0$ (no linear relationship)
  - $H_1 : \beta_j \neq 0$ (linear relationship does exist between $X_j$ and $Y$)
- Test statistic:

$$t = \frac{b_j - 0}{S_{b_j}}$$

$(\text{df} = n - k - 1)$

# Excel Output

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

**$t$ Stat for Price is $t_{STAT} = -2.306$, with $p$-value .0398.**

**$t$ Stat for Advertising is $t_{STAT} = 2.855$, with $p$-value .0145.**



| ANOVA | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 |
| Residual | 12 | 27033.306 | 2252.776 | | |
| Total | 14 | 56493.333 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

Curtin University

50 YEARS OF INNOVATION

# Example: Inferences About the Slope (1 of 2)

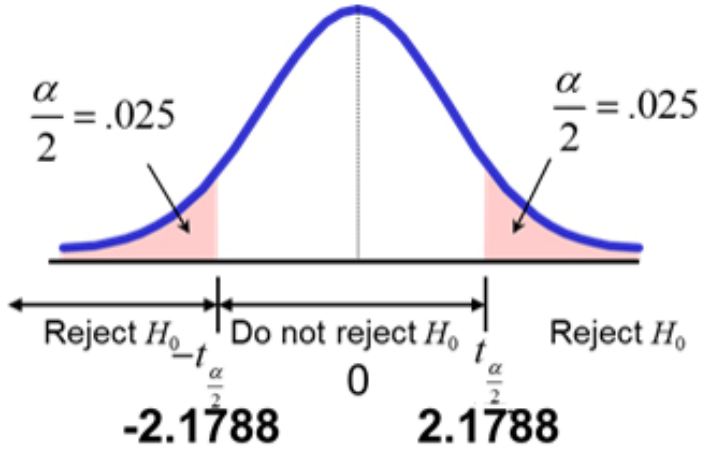
- $H_0 : \beta_j = 0$
  $H_1 : \beta_j \neq 0$
- $\text{df} = 15 - 2 - 1 = 12$, $\alpha = 0.05$
- $t_{\frac{\alpha}{2}} = 2.1788$
- From Excel output:

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| **Price** | **-24.97509** | **10.83213** | **-2.30565** | **0.03979** |
| **Advertising** | **74.13096** | **25.96732** | **2.85478** | **0.01449** |

Curtin University

# Example: Inferences About the Slope (2 of 2)

- The test statistic for each variable falls in the rejection region ($p$-values $< 0.05$).
- Decision: Reject $H_0$ for each variable.
- Conclusion: There is evidence that both Price and Advertising affect pie sales at $\alpha = 0.05$.



# Confidence Interval Estimate for the Slope (1 of 3)

- Confidence interval for the population slope $\beta_j$:

$$b_j \pm t_{\frac{\alpha}{2}} S_{b_j}$$

where $t$ has $(n - k - 1)$ d.f.

| | Coefficients | Standard Error |
|---|---|---|
| Intercept | 306.52619 | 114.25389 |
| Price | −24.97509 | 10.83213 |
| Advertising | 74.13096 | 25.96732 |



**Example:** Form a 95% confidence interval for the effect of changes in price $(X_1)$ on pie sales.

# Confidence Interval Estimate for the Slope (2 of 3)

- $-24.975 \pm (2.1788)(10.832)$
- So the interval is $(-48.576, -1.374)$
- This interval does not contain zero, so price has a significant effect on sales.

# Confidence Interval Estimate for the Slope (3 of 3)

| | *Coefficients* | *Standard Error* | ... | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|
| Intercept | 306.52619 | 114.25389 | ... | 57.58835 | 555.46404 |
| Price | −24.97509 | 10.83213 | ... | −48.57626 | −1.37392 |
| Advertising | 74.13096 | 25.96732 | ... | 17.55303 | 130.70888 |



- Excel output also reports these interval endpoints:
  - Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price, holding the effect of advertising constant.

# Using Dummy Variables

- A dummy variable is a categorical explanatory variable with two levels:
  - yes or no, on or off, male or female
  - coded as 0 or 1
- Assumes the slopes associated with numerical independent variables do not change with the value for the categorical variable.
- If more than two levels, the number of dummy variables needed is number of levels minus 1.

# Example: Dummy Variables with 2 Levels (1 of 2)

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

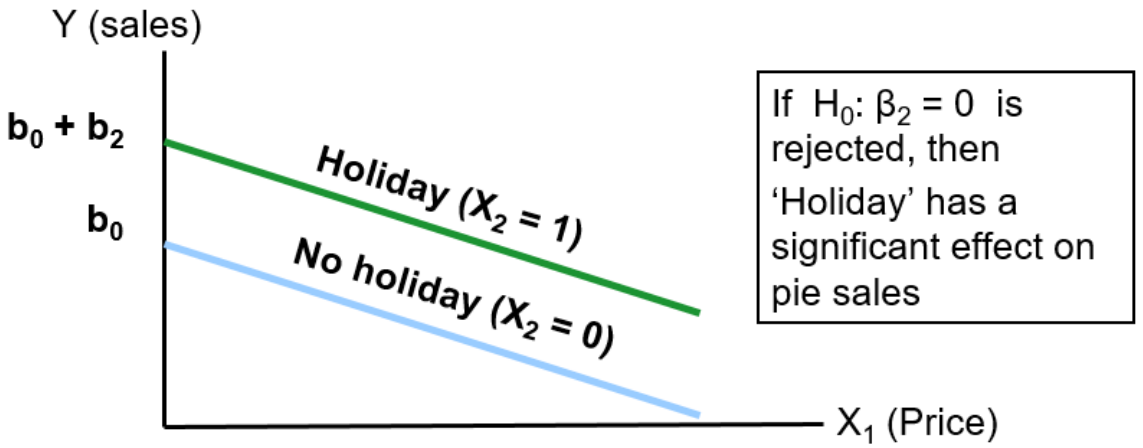
- Let $Y =$ pie sales, $X_1 =$ price
- $X_2 =$ holiday
  - $X_2 = 1$ if a holiday occurred during the week
  - $X_2 = 0$ if there was no holiday that week

# Example: Dummy Variables with 2 Levels (2 of 2)

$$\hat{Y} = b_0 + b_1 X_1 + b_2(1) = \boxed{(b_0 + b_2)} + \boxed{b_1 X_1} \quad \textbf{Holiday}$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2(0) = \boxed{b_0} + \boxed{b_1 X_1} \quad \textbf{No Holiday}$$

**Different intercept** **Same slope**




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

Curtin University

# Interpreting the Dummy Variable Coefficient (With 2 Levels)

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

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

# Interaction Between Independent Variables

- Hypothesizes interaction between pairs of $X$ variables.
  - Response to one $X$ variable may vary at different levels of another $X$ variable.
- Contains two-way cross product terms.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

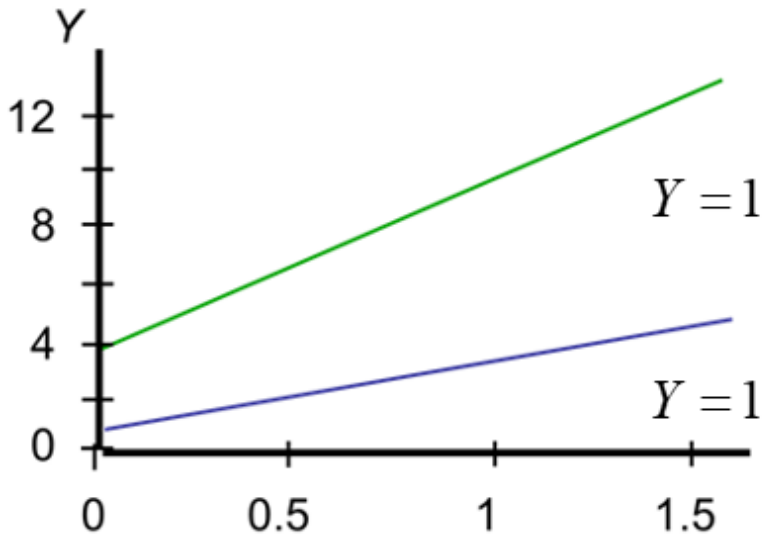
$$= b_0 + b_1 X_1 + b_2 X_2 + b_3 (X_1 X_2)$$

# Effect of Interaction

- Given: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- Without interaction term, effect of $X_1$ on $Y$ is measured by $\beta_1$.
- With interaction term, effect of $X_1$ on $Y$ is measured by $\beta_1 + \beta_3 X_2$.
- Effect changes as $X_2$ changes.

# Interaction Example

Suppose $X_2$ is a dummy variable and the estimated regression equation is: $\hat{Y} = 1 + 2X_1 + 3X_2 + 4X_1X_2$




$$X_2 = 1:$$
$$Y = 1 + 2X_1 + 3(1) + 4X_1(1) = 4 + 6X_1$$

$$X_2 = 0:$$
$$Y = 1 + 2X_1 + 3(0) + 4X_1(0) = 1 + 2X_1$$

Slopes are different if the effect of $X_1$ on $Y$ depends on $X_2$ value.

# Collinearity

- **Collinearity:** High correlation exists among two or more independent variables.
- This means the correlated variables contribute redundant information to the multiple regression model.
- Including two highly correlated independent variables can adversely affect the regression results.
- No new information provided:
  - Can lead to unstable coefficients (large standard error and low $t$-values)
  - Coefficient signs may not match prior expectations

# Some Indications of Strong Collinearity

- Incorrect signs on the coefficients.
- Large change in the value of a previous coefficient when a new variable is added to the model.
- A previously significant variable becomes non-significant when a new independent variable is added.
- The estimate of the standard deviation of the model increases when a variable is added to the model.