MODULE THREE: DETERMINING CAUSE AND MAKING RELIABLE FORECASTS

TOPIC 9: INTRODUCTION TO MULTIPLE REGRESSION

Learning Objectives

At the completion of this topic, you should be able to:

- construct a multiple regression model and analyse model output
- differentiate between independent variables and decide which ones to include in the regression model, and determine which independent variables are more important in predicting a dependent variable
- incorporate categorical variables in regression model
- detect collinearity

+The Multiple Regression Model

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

Multiple Regression Model with k Independent Variables:

Y-intercept

Population slopes

Random Error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

+Multiple Regression Equation

$$Y_i = \beta_0 + \boxed{\beta_1} X_{1i} + \boxed{\beta_2} X_{2i} + \ldots + \boxed{\beta_k} X_{ki} + \varepsilon_i$$

The <u>estimated</u> Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y

Estimated intercept

Estimated slope coefficients

$$\hat{Y}_i = \boxed{b_0} + \boxed{b_1} X_{1i} + \boxed{b_2} X_{2i} + \ldots + \boxed{b_k} X_{ki}$$

In this topic we will use Excel to obtain the regression slope coefficients and other regression summary measures

+Pie Sales Example:

Week	Pie Sales (Y)	Price (\$) (X_1)	Advertising (\$100s) (X_2)
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

A distributor of frozen dessert pies wants to evaluate factors thought to influence demand

Dependent variable:

Pie sales (units per week)

Independent variables:

Advertising ($100s), Price (in $)

Data are collected for 15 weeks

Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1 \text{ (Price \$)} + b_2 \text{ (Advertising \$100)}$$

+Multiple Regression 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	14730.01	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 (X_1)	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising (X_2)	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

+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 (**Y**)
- Price is in \$ (**X_1**)
- Advertising is in \$100s (**X_2**)

$Y = b_{01} + b_{11} X_1$
$Y = b_{02} + b_{12} X_2$
$Y = b_0 + b_1 X_1 + b_2 X_2$

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

+Using The Equation to Make Predictions

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

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

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

Predicted sales is 428.62 pies

Week	Pie Sales (Y)	Price (\$) (X_1)	Advertising (\$100s) (X_2)
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

+Coefficient of Multiple Determination (R Square/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 (Cont)

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

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

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

+Adjusted r^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?

$$\text{SE of the Estimate} = \sqrt{\frac{26{,}993.33}{11}} = 49.54$$

ANOVA	df		SS		MS	F	Significance F
Regression	2	+1 =3	29460.027	29,500	730.01	6.53861	0.01201
Residual	12	-1 =11	27033.306	26,993.33	'6		
Total	14		56493.333				

r^2 =29,500/56,493.33 = 52.22%

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

$$\text{Standard Error of the Estimate} = \sqrt{\frac{SSE}{Residual - DF}} = \sqrt{\frac{27033.306}{12}} = 47.46$$

+Adjusted r^2 (Cont)

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

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

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

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

$$r^2 = 1 - \frac{\text{Error sum of squares}}{\text{Total sum of squares}} = 1 - \frac{SSE}{SST}$$

$$r_{adj}^2 = 1 - \frac{\text{Mean Square Error}}{\text{Mean Square Total}} = 1 - \frac{\text{Error sum of squares/df}}{\text{Total sum of squares/df}} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

+Adjusted r^2 (Cont)

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

$r^2_{adj} = .44172$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

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

+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

Hypotheses:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)

H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

+F Test for Overall Significance

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			

Test statistic

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

where F has:

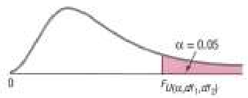
(numerator) = k, and

(denominator) = (n – k - 1) degrees of freedom

Table E.5

Critical values of *F*

For a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of *F* corresponding to a specified upper-tail area (α).



Denominator df_2	Numerator, df_1: 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90	243.90	245.90	248.00	249.10	250.10	251.10	252.20	253.30	254.30
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96

F Test for Overall Significance (Cont)

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

$$F = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.01	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

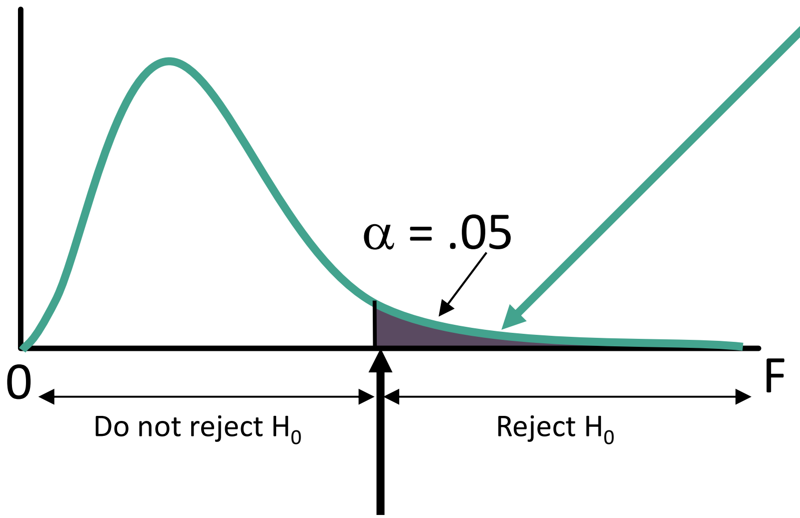
+F Test for Overall Significance (Cont)

H_0: $\beta_1 = \beta_2 = 0$

H_1: β_1 and β_2 not both zero

$\alpha = .05$

$df_1 = 2$ $df_2 = 12$



Critical Value: $F_\alpha = 3.885$ (Table E.5)

Test Statistic:

$$F = \frac{MSR}{MSE} = 6.5386$$

Decision:
Since F test statistic is in the rejection region (p-value < .05), reject H_0

Conclusion:
There is evidence that at least one independent variable affects Y

+Are Individual Variables Significant?

Shows if there is a linear relationship between the variable X_j and Y

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Hypotheses:

H_0: $\beta_j = 0$ (no linear relationship)

H_1: $\beta_j \neq 0$ (linear relationship does exist)

Use t tests of <u>individual</u> variable slopes (between X_j and Y)

$$\hat{Y}_i = b_0 + \boxed{b_1} X_{1i} + \boxed{b_2} X_{2i} + \ldots + \boxed{b_k} X_{ki}$$

+Are Individual Variables Significant? (Cont)

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 = -2.306, with p-value .0398

t-stat for Advertising is: t = 2.855, with p-value .0145

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

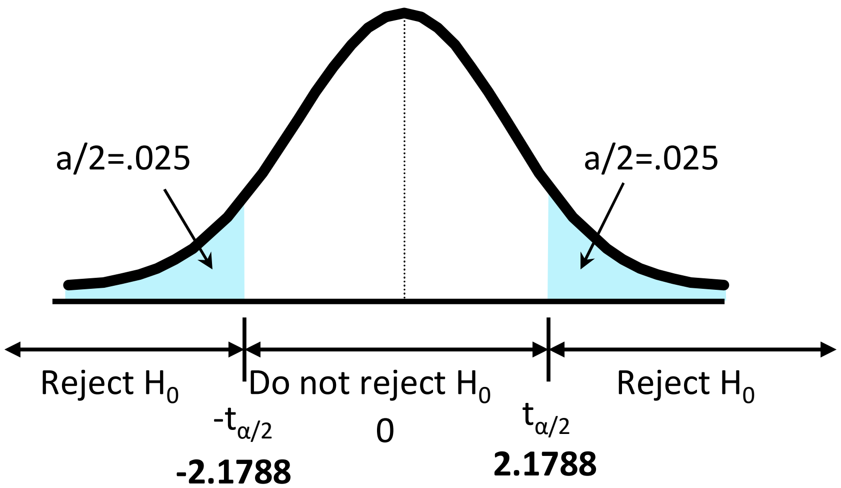
+Are Individual Variables Significant? (Cont)

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

d.f. = 15-2-1 = 12

α = .05 $t_{\alpha/2}$ = 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**

Decision:

The test statistic for each variable falls in the rejection region (p-values < .05)

Conclusion:

Reject H_0 for each variable.

There is evidence that both Price and Advertising affect pie sales at α = .05

+ Confidence Interval Estimate for the Slope

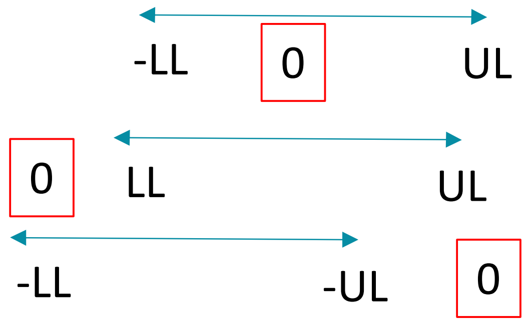
$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

Confidence interval for the population slope β_j

$$b_j \pm t_{n-k-1} S_{b_j}$$

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**

Here, t has: (15 – 2 – 1) = 12 d.f.

t = 2.1788

Example: Form a 95% confidence interval for the effect of changes in price (X_1) on pie sales: -24.975 ± (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 (Cont)

Confidence interval for the population slope β_i

	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**

Example: Excel output also reports these interval endpoints:

With 95% confidence, weekly sales are estimated to be reduced by between 1.37 to 48.58 pies on average for each increase of $1 in the selling price (assuming no change in the Advertising)

+ Using Dummy Variables

Examples: Gender/Department → Productivity/Job Satisfaction
Training course → Salesperson's performance

A dummy variable is a categorical explanatory variable with two levels:

- yes or no, on or off, male or other
- coded 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 minus 1

e.g. Department: Admin, Production, Distribution > 3 levels > 2 Dummy variables

+Dummy Variable Example (with 2 Levels):

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

Let: Y = pie sales

X_1 = price (Numerical variable)

X_2 = holiday (categorical variable)

(X_2 = 1 if a holiday occurred during the week)

(X_2 = 0 if there was no holiday that week)

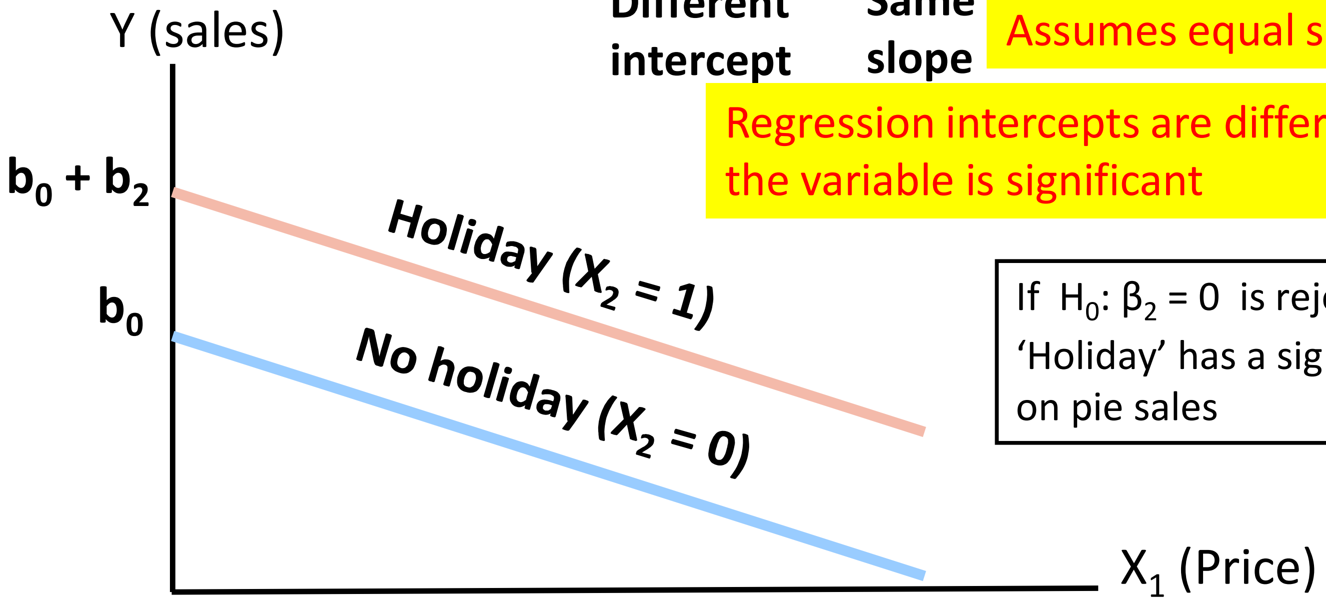
+Dummy Variable Example (with 2 Levels):

$\hat{Y} = b_0 + b_1X_1 + b_2(1) = \boxed{(b_0 + b_2)} + \boxed{b_1X_1}$	Holiday
$\hat{Y} = b_0 + b_1X_1 + b_2(0) = \boxed{b_0} + \boxed{b_1X_1}$	No holiday

Different intercept **Same slope**

Assumes equal slopes for other variables

Regression intercepts are different if the variable is significant



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

+Interpreting the Dummy Variable Coefficient - with 2 Levels

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

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

+Dummy Variable Models - more than 2 Levels

The number of dummy variables is **one less than the number of levels**

Example:

Y = apartment price ($000)

X_1 = size of apartment in hundreds of square metres

If number of bedrooms is incorporated:

Bedrooms = one, two, three (Number of levels =3)

Three levels, so two dummy variables are needed

+Dummy Variable Models - more than 2 Levels (Cont)

Example:

Let '1-bedroom' be the default category, and let X_2 and X_3 be used for the other two categories

Apartment price	Apartment size	Number of Bedrooms	X2	X3
XXXXX	XXX	1	0	0
XXXXX	XXX	2	1	0
		3	0	1

Y = apartment price

X_1 = size in hundreds of square metres

X_2 = 2 bedroom, 0 otherwise

X_3 = 3 bedroom, 0 otherwise

The estimated multiple regression equation is:

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

+Dummy Variable Models - more than 2 Levels (Cont)

Dummy Variables

Consider the regression equation:

$$\hat{Y} = 20.43 + 0.045X_1 + 18.84X_2 + 33.53X_3$$

For 1-bedroom: $X_2 = X_3 = 0$

$$\hat{Y} = 20.43 + 0.045X_1$$

For 2-bedroom: $X_2 = 1$; $X_3 = 0$

$$\hat{Y} = 20.43 + 0.045X_1 + \boxed{18.84}$$

With the same size in hundreds of square meters, a 2-bedroom will have an estimated average price of **18.84** thousand dollars more than a 1-bedroom apartment

For 3-bedroom: $X_2 = 0$; $X_3 = 1$

$$\hat{Y} = 20.43 + 0.045X_1 + \boxed{33.53}$$

With the same size in hundreds of square meters, a 3-bedroom will have an estimated average price of **33.53** thousand dollars more than a 1-bedroom apartment

+Collinearity (Multi-collinearity)

High correlation exists among two or more independent variables

Example: IVs – Height, Weight; DV – Pulse rate

Example: IVs – Income, Tax ; DV – Amount spent on groceries

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 error of the model increases when a variable is added to the model

To detect Collinearity – check correlations between IVs

– check the coefficients and/or *p*-values

+Measuring Collinearity Variance Inflationary Factor

The variance inflationary factor VIF_j can be used to measure collinearity:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where: R^2_j is the coefficient of multiple determination of independent variable X_j DV with all other X variables IVs

If: $VIF_j = 1$, X_j is uncorrelated with the other Xs

X_1 and X_2; X_1 and X_3 ; X_1 and X_2 ...

If: $VIF_j > 10$, X_j is highly correlated with the other Xs (conservative estimate reduces this to $VIF_j > 5$)