STAT5102

Regression in Practice

III. Multiple Linear Regression

Department of Statistics
The Chinese University of Hong Kong

Outline

In this chapter, we shall cover:

- First-Order model
- General linear regression model
  - The model in matrix terms
  - Estimation of regression coefficients
  - Fitted values and residuals
  - ANOVA
  - Inference about regression parameters
  - Coefficient of Multiple Determination
  - Estimation of mean response
  - Prediction of new observation
  - Confidence region
- Scatter Plot and residual plots
- Extra Sums of Squares
- Coefficients of Partial Determination
- Standardized multiple regression model
- Multicollinearity and its effects
- Polynomial Regression Models
- Interaction Regression Models
- Qualitative Predictors

First-Order Model with Two Predictors

Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

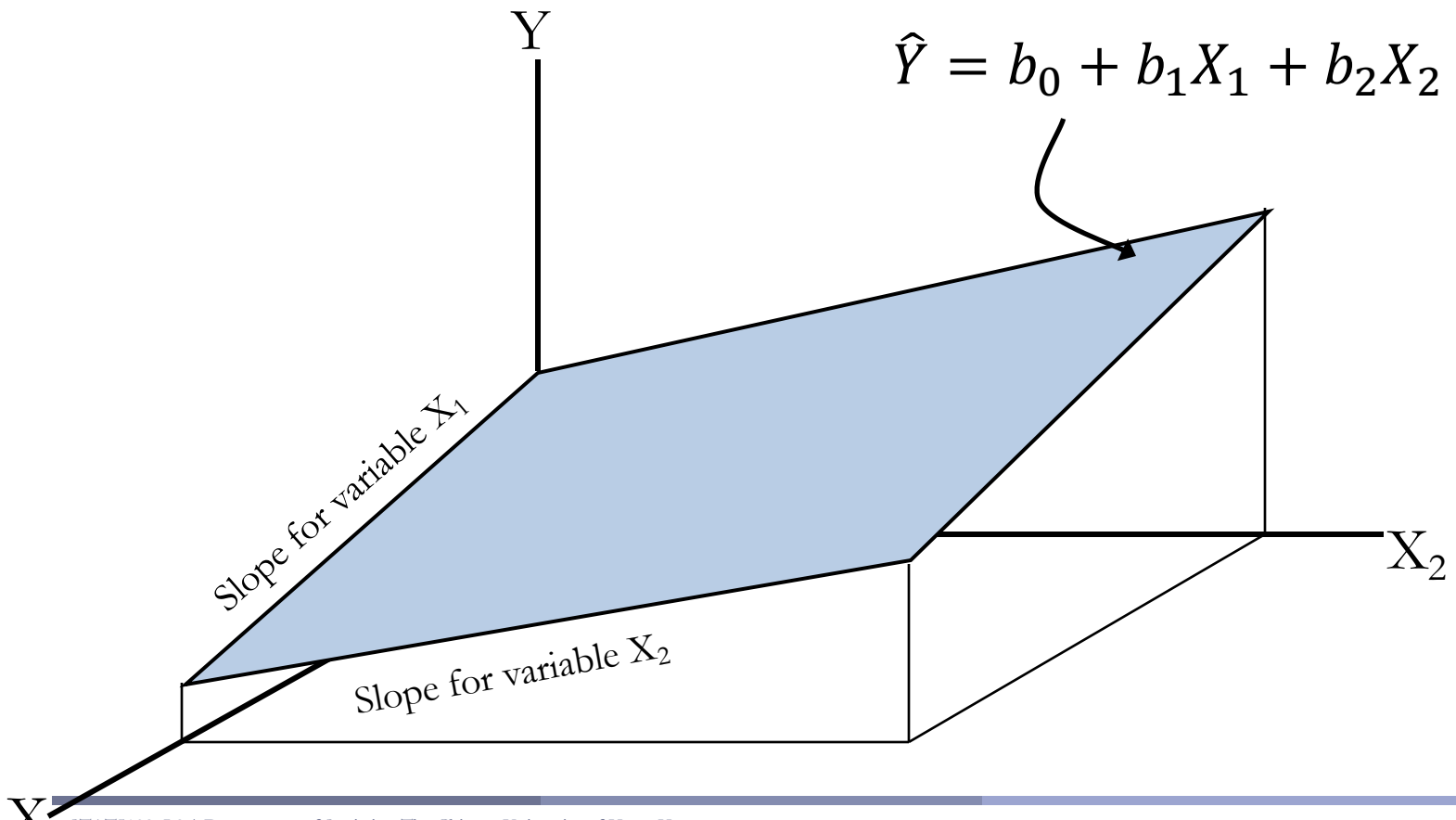
Assumption:

$$E(\varepsilon_i) = 0$$

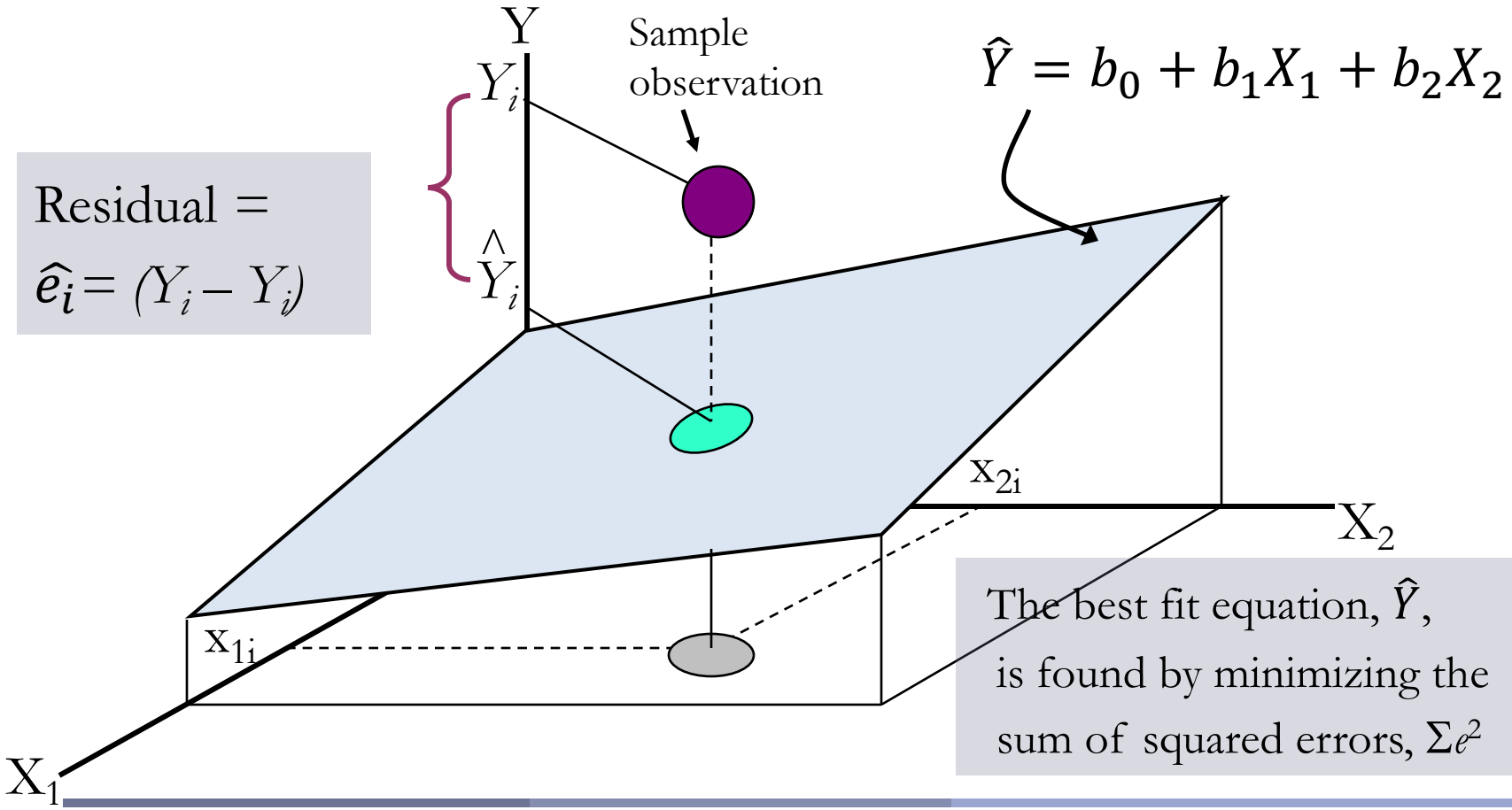
So,

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Response Surface (Plane)

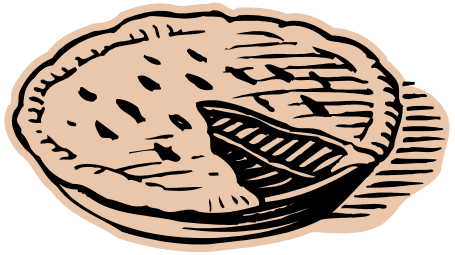


The Regression Equation



Example (Pie sales)

- 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 Example (cont'd)

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:

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

SAS output

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	306.52619	114.25389	2.68	0.0199
X2	Advertising	**1**	74.13096	25.96732	2.85	0.0145
X1	Price	**1**	-24.97509	10.83213	-2.31	0.0398

Regression Equation and its Interpretation

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

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

Regression Equation and its Interpretation

- The parameters β_1 and β_2 are sometimes called *partial regression coefficients* because they reflect the partial effect of one predictor variable when the other predictor variable is included in the model and is held constant.
- The two predictor variables are said to have *additive effects* or *not to interact*.

First-order Model with >2 Explanatory Variables

Model (k different predictor variables) :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Assumption:

$$E(\varepsilon_i) = 0$$

So,

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Multiple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

- $(X_1, \ldots, X_k)$ need not represent different predictor variables
- If the variables $X_1, \ldots, X_k$ represent k different predictor variables, the general linear regression model is a first-order model in which there are no interaction effects between the predictor variables.
- Qualitative predictor variables Vs Quantitative predictor variables. Eg. Gender [Indicator variables]
- Polynomial regression: contain squared and higher –order terms of the predictor variables.
- Transformed variables
- Interaction effects
- Combination of cases

Multiple Linear Regression Model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathrm{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,k} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

Multiple Linear Regression Model

- Assumptions
- Expected value of $\boldsymbol{Y}$
- Estimation of regression coefficients
- Fitted values and residuals
- ANOVA table

Is the Model Significant?

- *F*-Test for Overall Significance of the Model
- Shows if there is a linear relation between all of the X variables considered together and Y.
- Use F test statistic
- 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

Test statistic:

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

where F has (numerator) = k and (denominator) = n – (k+1) degrees of freedom.

Coefficient of Multiple Determination

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

$$R^2 = r^2_{Y.12..k} = \frac{\text{SSR}}{\text{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Adjusted R^2 (R_a^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 (R_a^2)

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

$$R_a^2 = 1 - \left[(1 - R^2) \left(\frac{n-1}{n-(k+1)} \right) \right]$$

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

This statistic penalise excessive use of unimportant independent variables and it is always smaller than R^2.

It provides a fair measure (on the goodness of fit) for comparing among models.

Estimation of Response Variable's mean for given $\boldsymbol{X}$

Let $\boldsymbol{X}_h$ denote the level of $\mathbf{X}$ for which we wish to estimate the mean response [to be estimated by $\hat{Y}_h$]. By the model,

$$\hat{Y}_h = \boldsymbol{X}_h^T \boldsymbol{b}$$

Distribution of $\hat{Y}_h$:

$$\hat{Y}_h \sim N\{\boldsymbol{X}_h^T \boldsymbol{\beta}, \sigma^2 \boldsymbol{X}_h (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_h\}$$

Confidence Interval for $E(Y_h)$

Two-sided 100(1-α)% C.I. for $E(Y_h)$:

$$\left(\hat{Y}_h - t_{\frac{\alpha}{2},n-(k+1)} S\sqrt{X_h^T (X^T X)^{-1} X_h}\,,\right.$$
$$\left.\hat{Y}_h + t_{\frac{\alpha}{2},n-(k+1)} S\sqrt{X_h^T (X^T X)^{-1} X_h}\,\right)$$

Prediction of a New Observation $Y_{h(new)}$

Prediction of $Y_{h(new)}$ corresponding to a given level $\boldsymbol{X}$ of the predictor variable by $\hat{Y}_h$.

Distribution of $Y_{h(new)} - \hat{Y}_h$:

$$Y_{h(new)} - \hat{Y}_h \sim N\{0, \sigma^2(1 + \boldsymbol{X}_h^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_h)\}$$

Confidence Interval for $Y_{h(new)}$

Two-sided 100(1-α)% C.I. for $Y_{h(new)}$:

$$\left(\hat{Y}_h - t_{\frac{\alpha}{2},n-(k+1)} S\sqrt{1 + X_h^T (X^T X)^{-1} X_h}\,,\right.$$
$$\left.\hat{Y}_h + t_{\frac{\alpha}{2},n-(k+1)} S\sqrt{1 + X_h^T (X^T X)^{-1} X_h}\,\right)$$

Prediction of mean of m new observations

Let the mean of m new observations be $\bar{Y}_{h(new)}$ for given $\boldsymbol{X_h}$. It is estimated by $\hat{Y}_h$.

Distribution of $\bar{Y}_{h(new)} - \hat{Y}_h$:

$$\bar{Y}_{h(new)} - \hat{Y}_h \sim \mathrm{N}\{0, \sigma^2(1/n + \boldsymbol{X}_h^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_h)\}$$

Confidence Interval for $\bar{Y}_{h(new)}$

Two-sided 100(1-α)% C.I. for $\bar{Y}_{h(new)}$:

$$\left(\hat{Y}_h - t_{\frac{\alpha}{2},n-(k+1)} S \sqrt{\frac{1}{m} + X_h^T (X^T X)^{-1} X_h}\right.,$$

$$\left.\hat{Y}_h + t_{\frac{\alpha}{2},n-(k+1)} S \sqrt{\frac{1}{m} + X_h^T (X^T X)^{-1} X_h}\right)$$

Confidence Region for Regression Surface

100(1-α)% confidence region for the entire regression surface over all combinations of values of the X variables is

$$\left(\hat{Y}_h - Ws\{\hat{Y}_h\}, \quad \hat{Y}_h + Ws\{\hat{Y}_h\}\right)$$

where

$$W^2 = p\, F_{a;\,(k+1),n-(k+1)}$$

Scatter Plots and Residual Plots

- Scatter plot matrix
- Residual plots
- Normal probability plot

Example

Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialise in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2).

- Y: Sales in a community
- X_1: number of persons aged 16 or younger in the community
- X_2: per capita disposable personal income in the community

ANOVA Table

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	24015	12008	99.10	<.0001
Error	18	2180.92741	121.16263		
Corrected Total	20	26196			

H_0: $\beta_1 = \beta_2 = 0$ (no linear relation)
H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

Since the *p*-value of the test is less than .0001, reject the null hypothesis.

The Fitted Regression Model

$$\widehat{\text{Sales}} = -68.85707 + 1.45456\, X_1 + 9.3655\, X_2$$

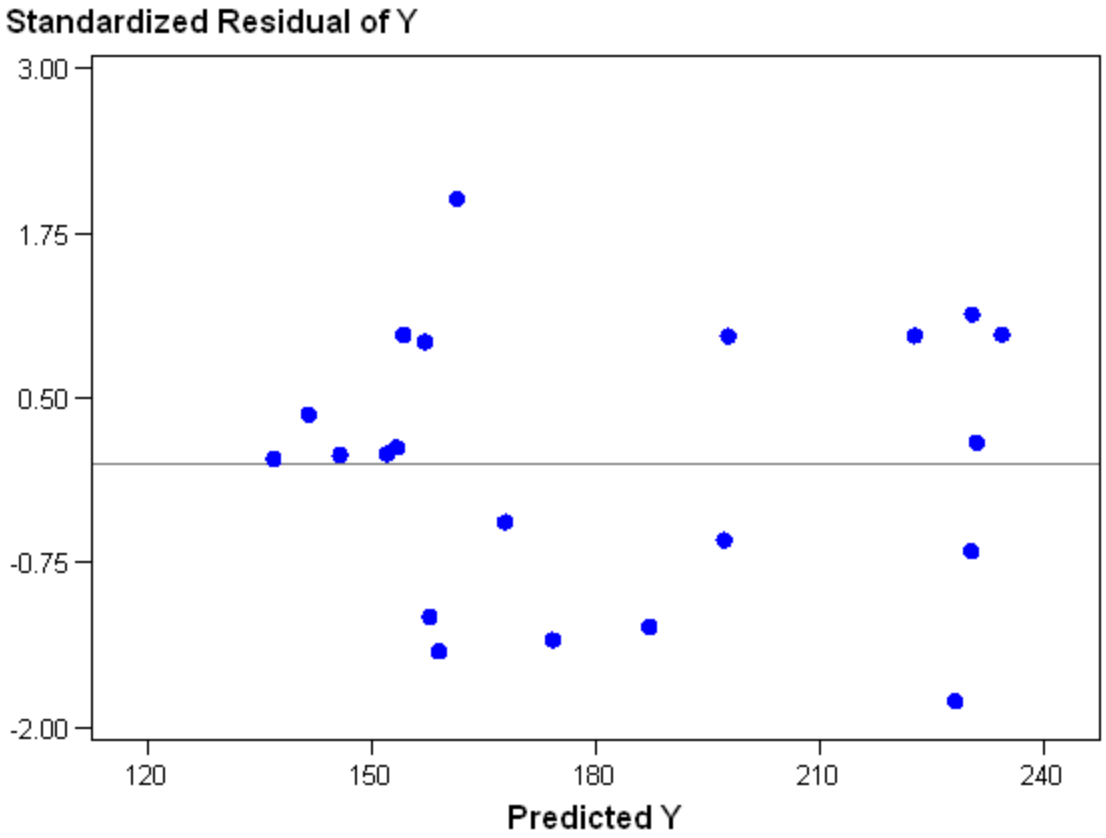
Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**
Intercept	Intercept	**1**	-68.85707	60.01695	-1.15	0.2663	0
X2	per capita disposable personal income	**1**	9.36550	4.06396	2.30	0.0333	0.25110
X1	number of persons aged 16 or younger	**1**	1.45456	0.21178	6.87	<.0001	0.74837

Estimate of the Covariance Matrix of the Estimates

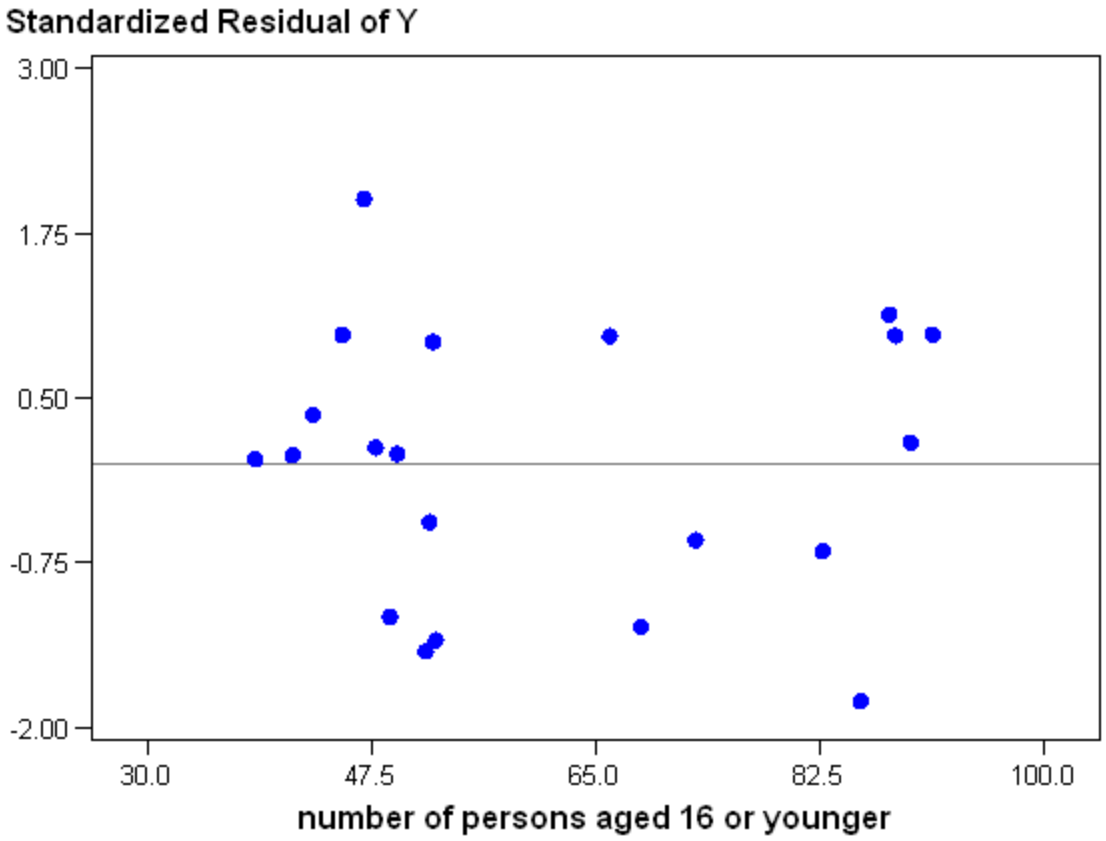
Covariance of Estimates				
Variable	**Label**	**Intercept**	**X2**	**X1**
Intercept	**Intercept**	3602.0346743	-241.4229923	8.7459395806
X2	**per capita disposable personal income**	-241.4229923	16.515755794	-0.672442604
X1	**number of persons aged 16 or younger**	8.7459395806	-0.672442604	0.0448515096

The use of the above covariance matrix will be discussed in class.

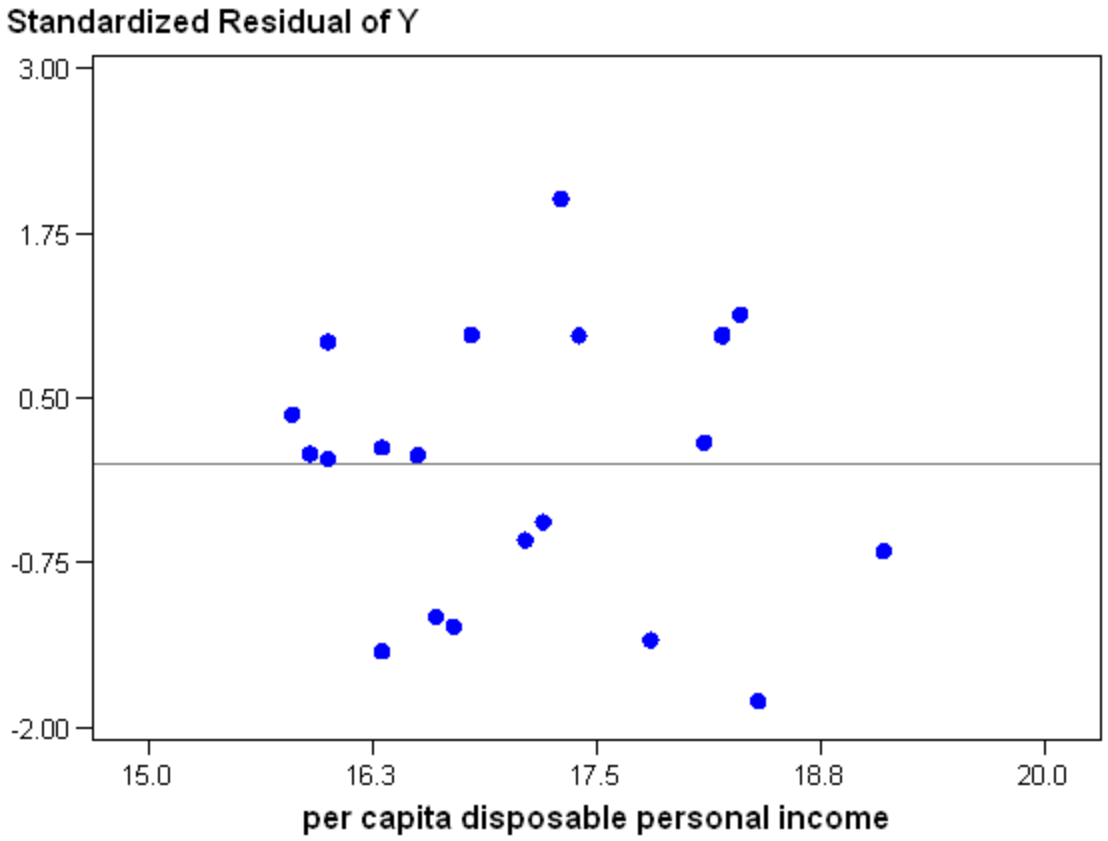
Residual Plot (vs Fitted Values)



Residual Plot (vs X_1)



Residual Plot (vs X_2)



R^2 and adjusted R^2 (R_a^2)

Root MSE	11.00739	**R-Square**	0.9167
Dependent Mean	181.90476	**Adj R-Sq**	0.9075
Coeff Var	6.05118		

Extra Sum of Squares

A measurement of the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model.

Extra Sum of Squares

Contribution of a Single Independent Variable X_j

$$\begin{aligned}
&\text{SSR}(X_j \mid \text{all variables except } X_j) \\
&= \text{SSR (all variables)} - \text{SSR(all variables except } X_j) \\
&= \text{SSE (all variables except } X_j) - \text{SSE (all variables)}
\end{aligned}$$

Measures the contribution of X_j in explaining the total variation in Y (SST).

Extra Sum of Squares [eg X_1 and X_2]

Contribution of X_2

$$\begin{aligned} &SSR(X_2 \mid X_1) \\ &= SSR(X_1, X_2) - SSR(X_1) \\ &= SSE(X_1) - SSE(X_1, X_2) \end{aligned}$$

Measures the contribution of X_2 in explaining the total variation in Y (SST)

Example

Study of the relation of amount of body fat (Y) to several possible predictor variables, based on a sample of 20 healthy females 25 – 34 years old.

Y = Body Fat

X_1= Triceps skin fold thickness

X_2= Thigh circumference

X_3 = Mid-arm circumference

Regression of Y on X_1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	-1.49610	3.31923	-0.45	0.6576
X1	Triceps Skinfold Thickness	**1**	0.85719	0.12878	6.66	<.0001

Regression of Y on X_2

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	-23.63449	5.65741	-4.18	0.0006
X2	Thigh Circumference	**1**	0.85655	0.11002	7.79	<.0001

Regression of Y on X_1 and X_2

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	-19.17425	8.36064	-2.29	0.0348
X1	Triceps Skinfold Thickness	**1**	0.22235	0.30344	0.73	0.4737
X2	Thigh Circumference	**1**	0.65942	0.29119	2.26	0.0369

Regression of Y on X_1, X_2 and X_3

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	117.08469	99.78240	1.17	0.2578
X3	Midarm Circumference	**1**	-2.18606	1.59550	-1.37	0.1896
X1	Triceps Skinfold Thickness	**1**	4.33409	3.01551	1.44	0.1699
X2	Thigh Circumference	**1**	-2.85685	2.58202	-1.11	0.2849

Extra Sum of Squares

$$\text{SSR}(X_2 \mid X_1) = \text{SSE}(X_1) - \text{SSE}(X_1, X_2)$$

$$= 143.12 - 109.95 = 33.17$$

Or

$$\text{SSR}(X_2 \mid X_1) = \text{SSR}(X_1, X_2) - \text{SSR}(X_1)$$

$$= 385.44 - 352.27 = 33.17$$

Extra Sum of Squares

$$SSR(X_3 \mid X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$$

$$= 109.95 - 98.41 = 11.54$$

Or

$$SSR(X_3 \mid X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$$

$$= 396.98 - 385.44 = 11.54$$

Extra Sum of Squares

Marginal effect of adding several variables

$$SSR(X_2, X_3 | X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$$
$$= 143.12 - 98.41 = 44.71$$

Or

$$SSR(X_2, X_3 | X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$$
$$= 396.98 - 352.27 = 44.71$$

Decomposition of SSR into Extra Sum of Squares

Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Type I SS**
Intercept	Intercept	**1**	117.08469	99.78240	1.17	0.2578	8156.76050
X1	Triceps Skinfold Thickness	**1**	4.33409	3.01551	1.44	0.1699	352.26980
X2	Thigh Circumference	**1**	-2.85685	2.58202	-1.11	0.2849	33.16891
X3	Midarm Circumference	**1**	-2.18606	1.59550	-1.37	0.1896	11.54590

The Partial *F*-Test Statistic

H_0: $\beta_q = \beta_{q+1} = \cdots = \beta_k = 0$

H_1: not all of the β's in the null hypothesis equal zero

Test statistics: $F = \dfrac{SSE(R) - SSE(F)}{df_R - df_F} \div \dfrac{SSE(F)}{df_F}$

Decision rule: Reject H_o if $F \geq F_{\alpha, k+1-q, n-(k+1)}$

Special Cases

- Test whether all β's = 0
- Test whether a single $\beta = 0$
- Test whether some β's = 0

Example on Body Fat (cont'd)

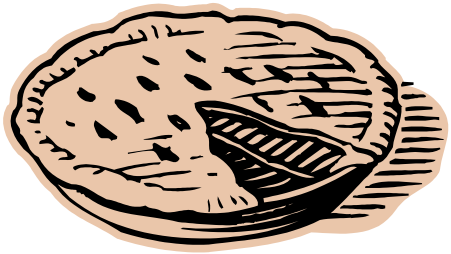
- Test whether X_3 can be dropped from the model

- Test whether X_2 and X_3 can be dropped from the full model.

Another Example (Pie Sales)

Example: Frozen desert pies

Test at the $\alpha = 0.05$ level to determine whether the price variable significantly improves the model given that advertising is included.



Another Example (Pie Sales)

H_0: X_1 (price) does not improve the model with X_2 (advertising) included

H_1: X_1 does improve model

$\alpha = .05$, d.f.'s = 1 and 12

F critical value = 4.75

Another Example (Pie Sales)

(For X_1 and X_2)

ANOVA			
	df	SS	MS
Regression	2	29460.02687	14730.01343
Residual	12	27033.30647	2252.775539
Total	14	56493.33333	

(For X_2 only)

ANOVA		
	df	SS
Regression	1	17484.22249
Residual	13	39009.11085
Total	14	56493.33333

$$F = \frac{\text{SSR}(X_1 \mid X_2)}{\text{MSE(all)}} = \frac{29{,}460.03 - 17{,}484.22}{2252.78} = 5.316$$

Conclusion: Reject H_0; adding X_1 does improve model

Coefficients of Partial Determination

$$R^2_{Yj.(\text{all variables except j})} = \frac{SSR(X_j \mid \text{all variables except j})}{SSE(\text{all variables except j})}$$

- Measures the proportion of variation in the dependent variable that is explained by X_j while controlling for (holding constant) the other explanatory variables
- Coefficients of partial correlation

Standardized Multiple Regression Model

- To control roundoff errors in normal equations calculations
- To permit comparisons of the estimated regression coefficients in common units
- No intercept term

Multicollinearity and its Effects

- High correlation exists among two or more independent variables.

- This means the correlated variables contribute redundant information to the multiple regression model.

Multicollinearity and its effects

Including two highly correlated explanatory 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

Multicollinearity and its effects

Some indications of strong multicollinearity:

- 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 insignificant 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

Detection of multicollinearity and remedial measures will be discussed later.

Polynomial Regression Models

When are polynomial regression models being used?

- When the true curvilinear response function is indeed a polynomial function

- When the true curvilinear response function is unknown (or complex) but a polynomial function is a good approximation to the true function.

Polynomial Regression Models

Example: 1 predictor variable, second order

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where

$$x_i = X_i - \overline{X}$$

The reason for using a centered predictor variable in the polynomial regression model is that X_1 and X_2 often will be highly correlated. Centering the predictor variable often reduces the multicollinearity substantially, and tends to avoid computational difficulties.

Polynomial Regression Models

Example: 2 predictor variables, second order

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

where

$$x_{i1} = X_{i1} - \overline{X}_1 \qquad x_{i2} = X_{i2} - \overline{X}_2$$

Interaction Regression Models

- 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

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

$$= b_0 + b_1X_1 + b_2X_2 + b_3(X_1X_2)$$

Interaction Regression Models

Example: 3 predictor variables

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \varepsilon_i$$

where

$$x_{i1} = X_{i1} - \overline{X}_1 \qquad x_{i2} = X_{i2} - \overline{X}_2 \qquad x_{i3} = X_{i3} - \overline{X}_3$$

Effect of Interaction

- Given: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

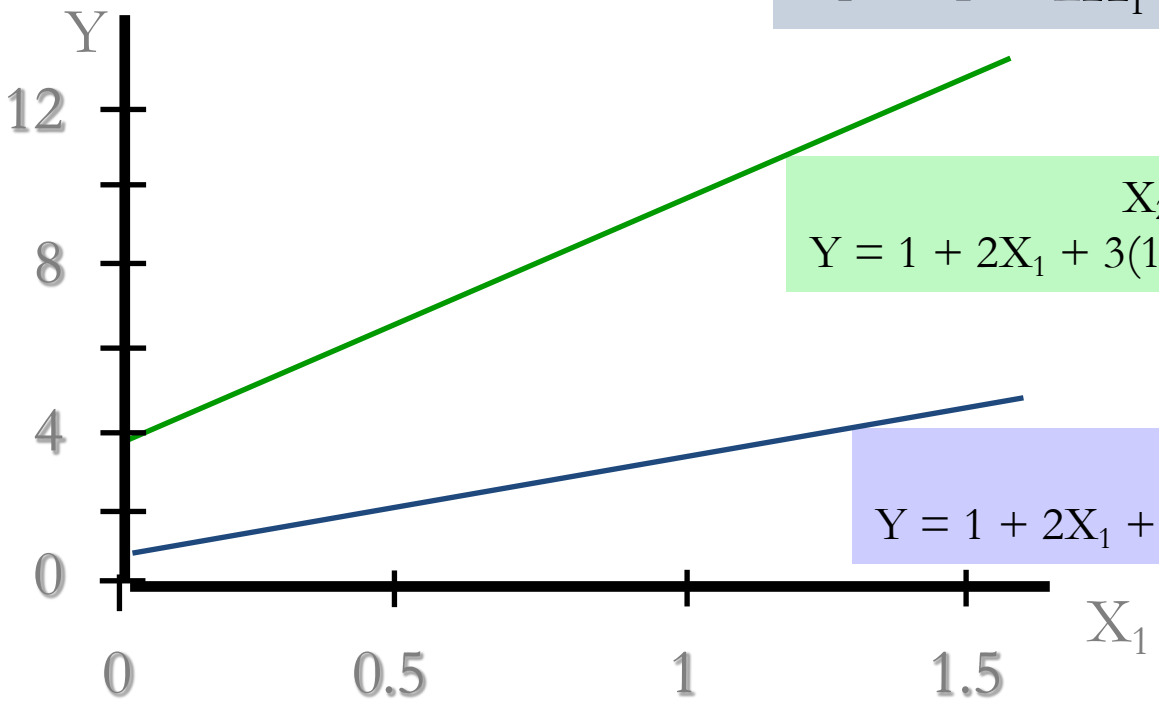
- Without interaction term, effect of X_1 on Y is measured by β_1
- With interaction term, effect of X_1 on Y is measured by $\beta_1 + \beta_3 X_2$

- Effect changes as X_2 changes

Effect of Interaction

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

Interaction Regression Models

- Additive model
- Reinforcement interaction effect
- Interference interaction effect

Significance of Interaction Term

- Can perform a partial *F*-test for the contribution of a variable to see if the addition of an interaction term improves the model

- Multiple interaction terms can be included

  Use a partial *F*-test for the simultaneous contribution of multiple variables to the model

Qualitative Predictors

- To quantify qualitative predictors, we use indicator variables (dummy variables).

- An indicator variable is a categorical explanatory variable with two levels:
  - yes or no, on or off, male or female
  - coded as 0 or 1

- If more than two levels, the number of indicator variables needed is (number of levels - 1)

Indicator-Variable Example (with 2 Levels)

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Let:

Y = pie sales

X_1 = price

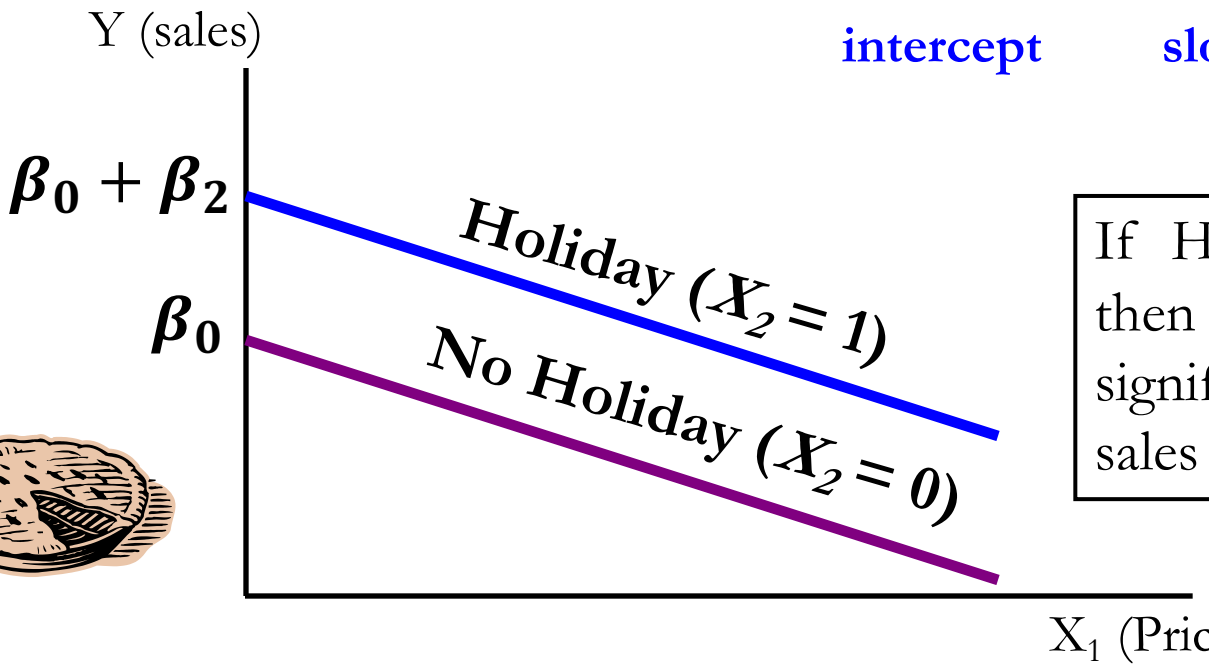
X_2 = holiday

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

Indicator-Variable Example (with 2 Levels)

$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_1$		**Holiday**
$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2(0) = (\beta_0) + \beta_1 X_1$		**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 Indicator Variable Coefficient (2 Levels)

Example: Sales = 300-30(Price)+15(Holiday)

Sales: number of pies sold per week

Price: pie price in $

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

$\beta_2 = 15$: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

Indicator-Variable Models (more than 2 Levels)

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

Y = house price; X_1 = square feet

If style of the house is also thought to matter:

Style = ranch, split level, condo

Three levels, so two dummy variables are needed

Indicator-Variable Models (more than 2 Levels)

Example: Let "condo" be the default category, and let X_2 and X_3 be used for the other two categories:

Y = house price
X_1 = square feet
X_2 = 1 if ranch, 0 otherwise
X_3 = 1 if split level, 0 otherwise

The multiple regression equation is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Interpreting the indicator variable coefficients (3 Levels)

Consider the regression equation:

$$\hat{Y} = 20.43 + 0.045X_1 + 23.53X_2 + 18.84X_3$$

For a condo: $X_2 = X_3 = 0$

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

For a ranch: $X_2 = 1; X_3 = 0$

$$\hat{Y} = 20.43 + 0.045X_1 + 23.53$$

With the same square feet, a ranch will have an estimated average price of 23.53 thousand dollars more than a condo

For a split level: $X_2 = 0; X_3 = 1$

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

With the same square feet, a split-level will have an estimated average price of 18.84 thousand dollars more than a condo.