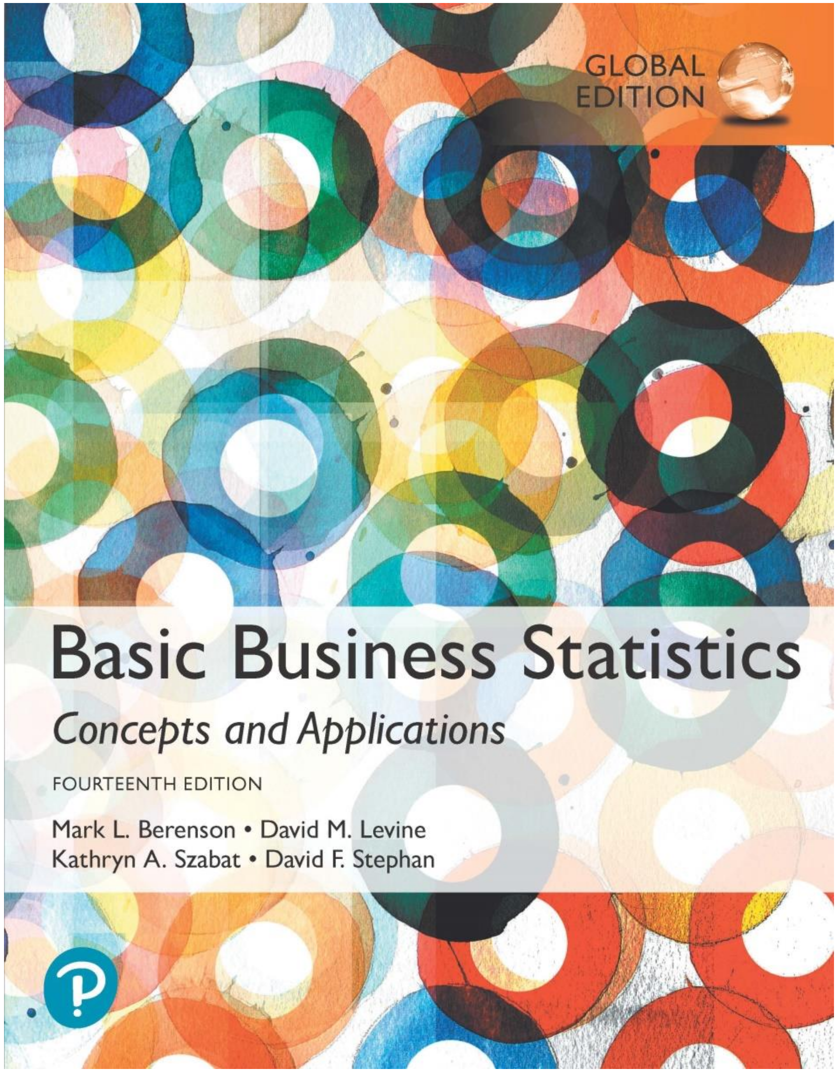


Chapter 14 & 15

Introduction to Multiple Regression

Objectives

In this chapter, you learn:

- How to develop a multiple regression model.
- How to interpret the regression coefficients.
- How to evaluate the assumptions in multiple regression
- How to determine which independent variables to include in the regression model.
- How to determine which independent variables are most important in predicting a dependent variable
- How to build a multiple regression model

The Multiple Regression Model With k Independent Variables

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

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

Where:

β_0 = Y intercept
β_1 = slope of Y with variable X_1, holding $X_2, X_3, X_4, \ldots, X_k$ constant
β_2 = slope of Y with variable X_2, holding $X_1, X_3, X_4, \ldots, X_k$ constant
β_3 = slope of Y with variable X_3, holding $X_1, X_2, X_4, \ldots, X_k$ constant
.
.
β_k = slope of Y with variable X_k, holding $X_1, X_2, X_3, \ldots, X_{k-1}$ constant
ε_i = random error in Y for observation i

Multiple Regression Model With 2 Independent Variables

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

Where:

β_0 = Y intercept

β_1 = slope of Y with variable X_1, holding X_2 constant

β_2 = slope of Y with variable X_2, holding X_1 constant

ε_i = random error in Y for observation i

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

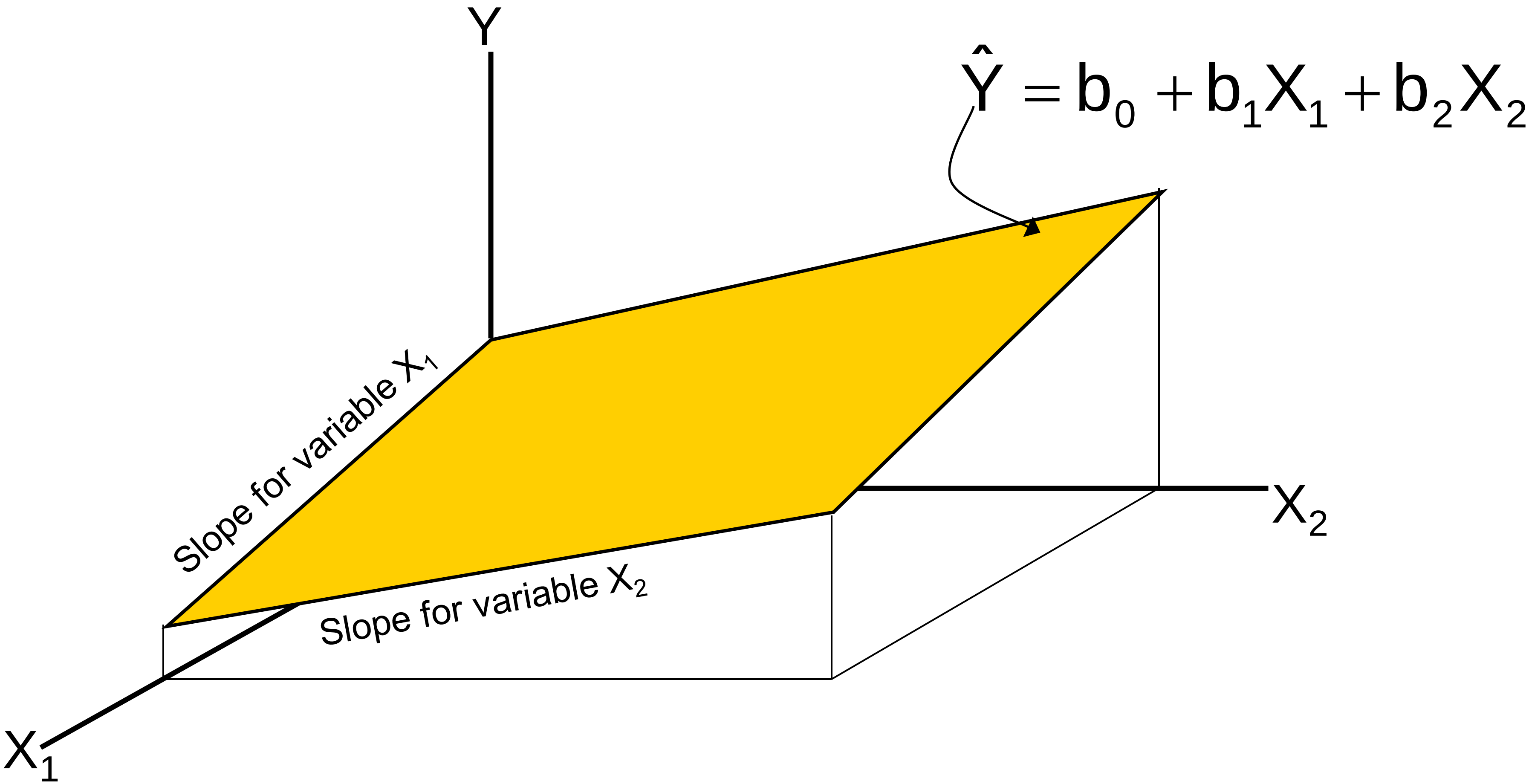
Estimated intercept

Estimated slope coefficients

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

Multiple Regression Equation With Two Independent Variables

Two variable model



Example:
2 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:

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

Excel 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**	**14730.013**	**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**

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 (=holding constant variable 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 (=holding constant variable price').

Using The Regression Equation to Make Predictions

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

$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.6216\end{aligned}$$

Predicted sales is 428.6216 pies.

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

Using The Regression Equation to Make Predictions

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

$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})\\ &= 306.526 - 24.975(5.50) + 74.131(3.5)\\ &= 428.6216\end{aligned}$$

Predicted sales is 428.6216 pies.

Careful: You should only predict <u>within the range</u> of the values of all the independent variables

Predictions in Excel

	A	B
1	**Confidence and Prediction Estimate Intervals**	
2		
3	**Data**	
4	**Confidence Level**	**95%**
5		
6	**Price given value**	**5.5**
7	**Advertising given value**	**3.5**
8		
20	t Statistic	2.178813
21	**Predicted Y (YHat)**	**428.6216**
22		
23	**For Average Predicted Y (YHat)**	
24	**Interval Half Width**	**37.50306**
25	**Confidence Interval Lower Limit**	**391.1185**
26	**Confidence Interval Upper Limit**	**466.1246**
27		
28	**For Individual Response Y**	
29	**Interval Half Width**	**110.0041**
30	**Prediction Interval Lower Limit**	**318.6174**
31	**Prediction Interval Upper Limit**	**538.6257**

Input values

Predicted $\hat{Y}$ value

Confidence interval for the mean value of Y, given these X values.

Prediction interval for an individual Y value, given these X values.

The Coefficient of Multiple Determination, r^2

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

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

Multiple Coefficient of Determination In Excel

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

$$r^2 = \frac{SSR}{SST} = \frac{29,460.027}{56,493.306} = .52148$$

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	14730.013	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

Adjusted r^2

- r^2 never decreases when a new X variable is added to the model.
  - This can be a disadvantage <u>when comparing models</u>.
- What is the net effect of adding a new variable?
  - Did the new X variable add explanatory power to the model?

Adjusted r^2

(continued)

- 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)

  - Penalizes excessive use of unimportant independent variables.
  - Smaller than r^2.
  - <u>Useful in comparing among models</u>.

Is the Overall 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: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (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:

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

where F_{STAT} has numerator d.f. = k and
denominator d.f. = (n – k - 1)

F Test for Overall Significance

(continued)

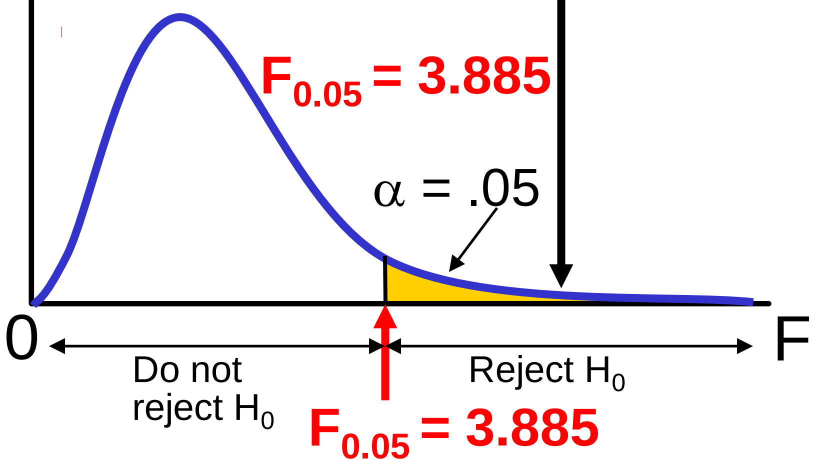
H_0: $\beta_1 = \beta_2 = 0$

H_1: β_1 and β_2 not both zero

$\alpha = .05$

$df_1 = 2 \qquad df_2 = 12$

Critical Value:



Test Statistic:

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

Decision:

Since F_{STAT} test statistic is in the rejection region (p-value < .05), reject H_0.

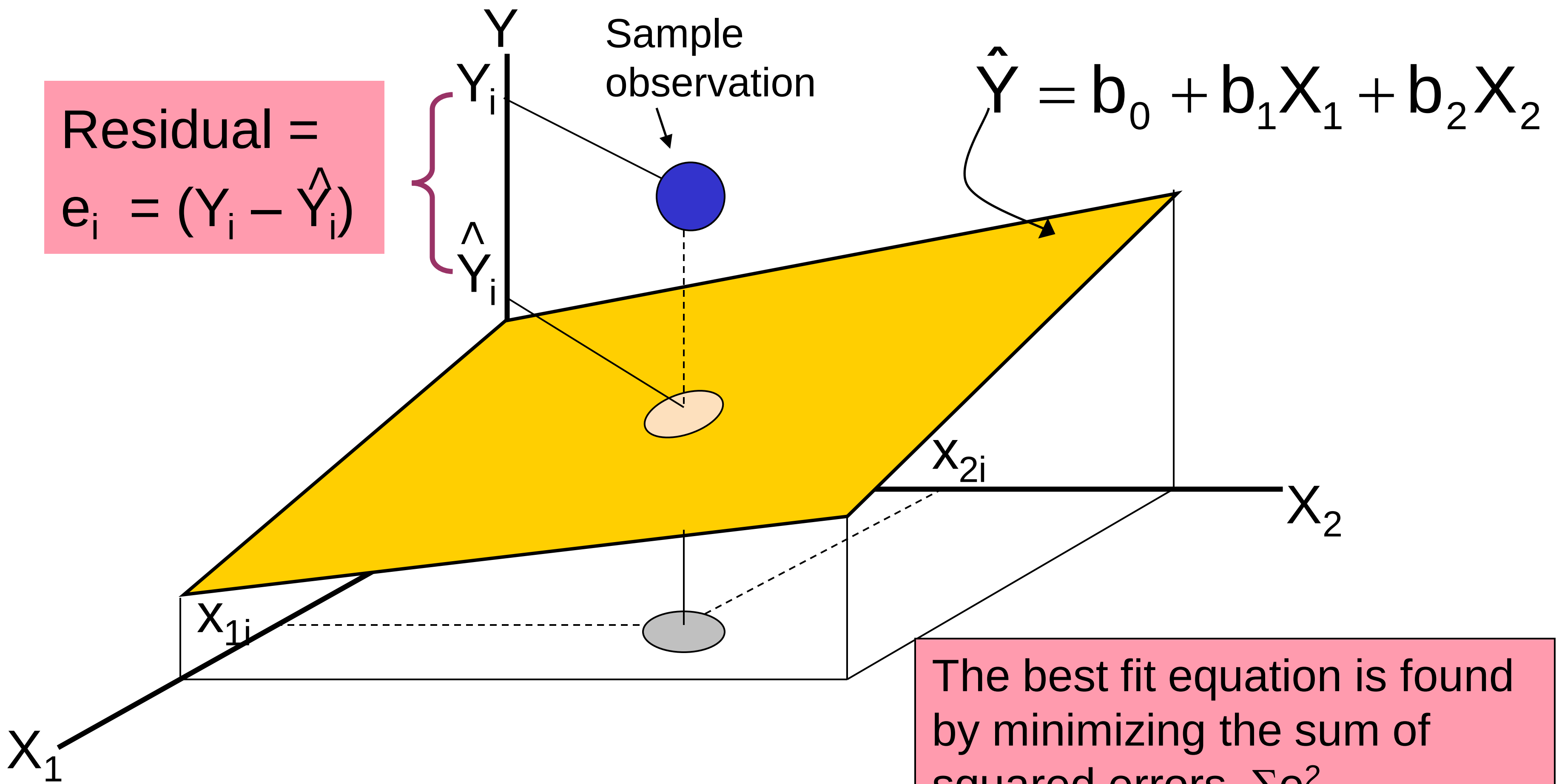
Conclusion:

There is evidence that <u>at least one independent variable</u> affects Y.

***Tip**: You can also use the p-value from Excel to reach the same conclusion*

Residuals in Multiple Regression

Two variable model



Residual =
$e_i = (Y_i - \hat{Y}_i)$

The best fit equation is found by minimizing the sum of squared errors, Σe^2.

Multiple Regression Assumptions

Errors (residuals) from the regression model:

$$\boxed{e_i = (Y_i - \hat{Y}_i)}$$

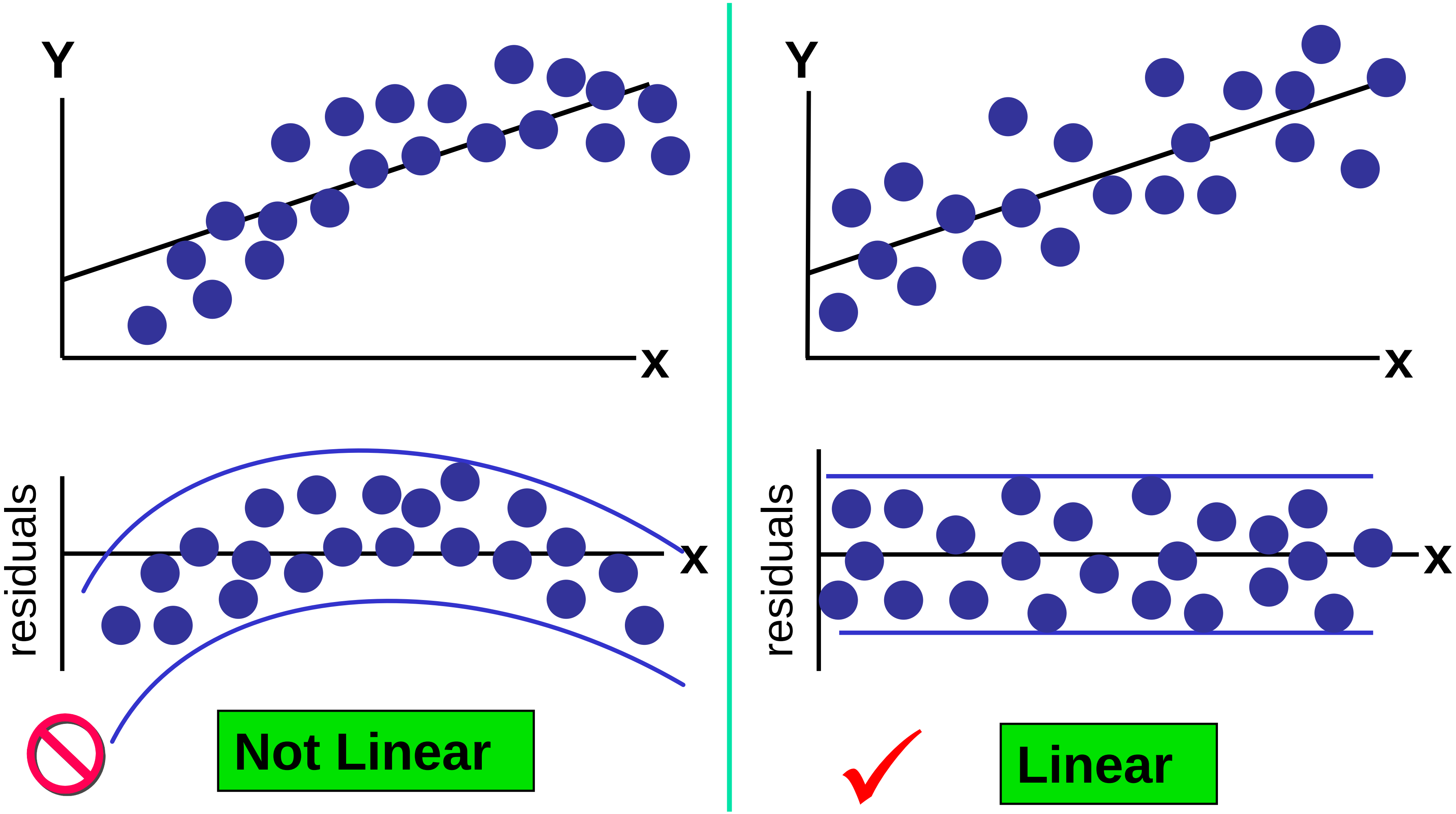
Assumptions:

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

Assumptions of Regression L.I.N.E

- Linearity:
  - The relationship between X and Y is linear.
- Independence of Errors (autocorrelation – Durbin Watson test):
  - Error values are statistically independent.
  - Particularly important when data are collected over a period of time (time series).
- Normality of Error:
  - Error values are normally distributed for any given value of X.
- Equal Variance (also called homoscedasticity) (White test):
  - The probability distribution of the errors has constant variance.

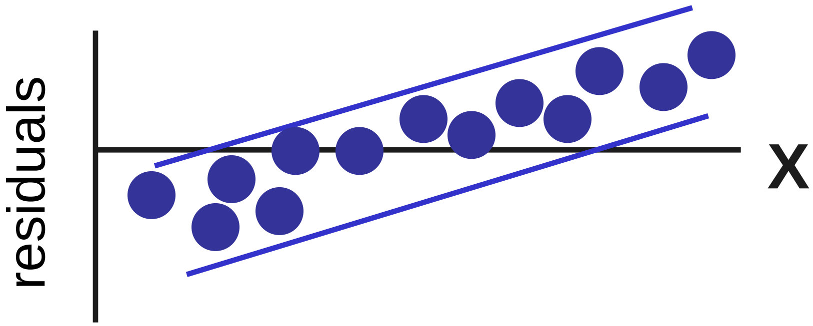
Residual Analysis for Linearity

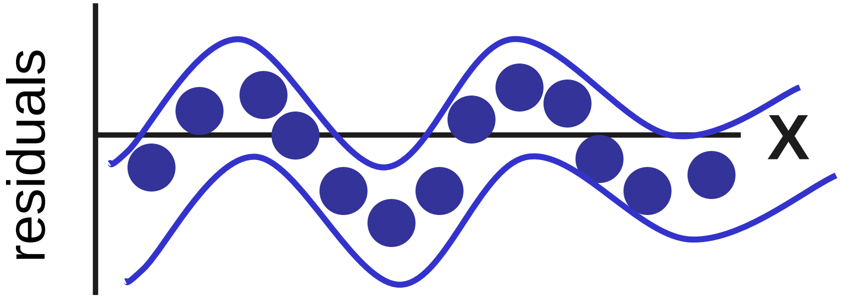


Residual Analysis for Independence (autocorrelation)

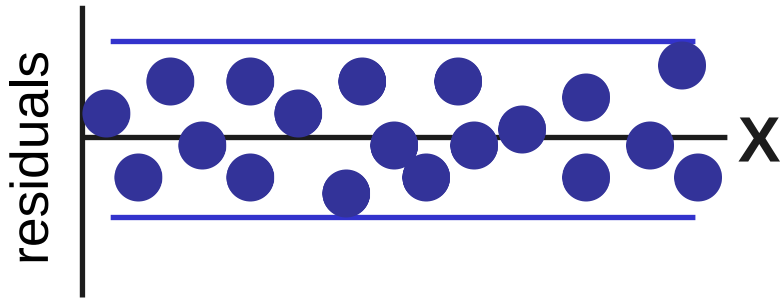
Cyclical Pattern:

Not Independent





No Cyclical Pattern
Independent



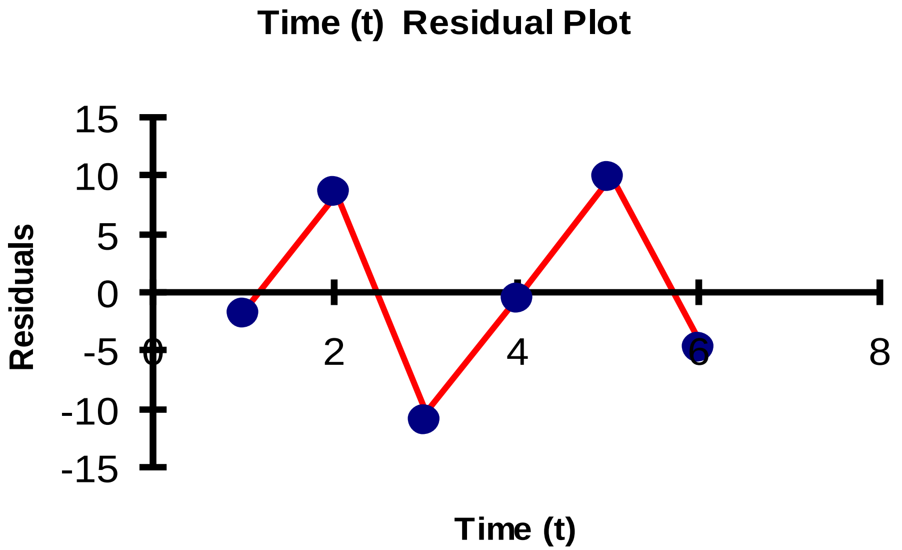
Measuring Autocorrelation: The Durbin-Watson Statistic

- Used when data are collected over time to detect if autocorrelation is present.
- Autocorrelation exists if residuals in one time period are related to residuals in another period.

Autocorrelation

- Autocorrelation is correlation of the errors (residuals) over time.

- Here, residuals show a cyclical pattern, not random. Cyclical patterns are a sign of positive autocorrelation.



- Violates the regression assumption that residuals are random and independent.

The Durbin-Watson Statistic

- The Durbin-Watson statistic is used to test for autocorrelation.

H_0: positive autocorrelation does not exist

H_1: positive autocorrelation is present

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- The possible range is $0 \le D \le 4$.
- D should be close to 2 if H_0 is true.
- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation.

Testing for Positive Autocorrelation

H_0: positive autocorrelation **does not exist**

H_1: positive autocorrelation is present

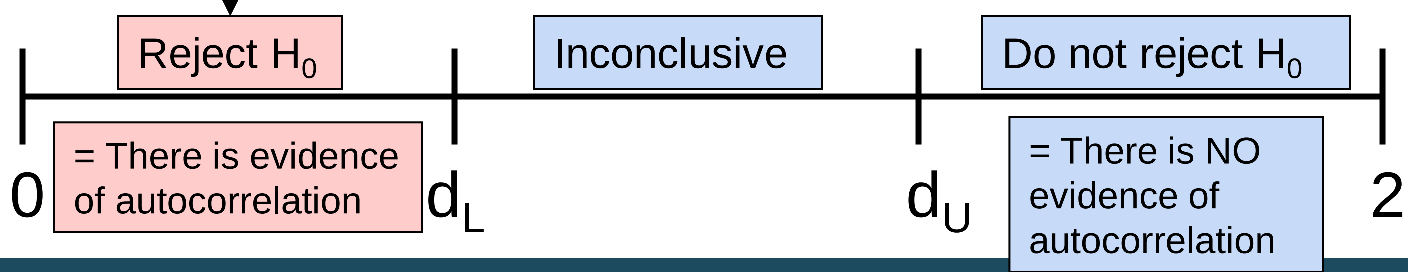
- Calculate the Durbin-Watson test statistic = D.

  (The Durbin-Watson Statistic can be found using Excel.)

- Find the values d_L and d_U from the Durbin-Watson table (E8 in your book).

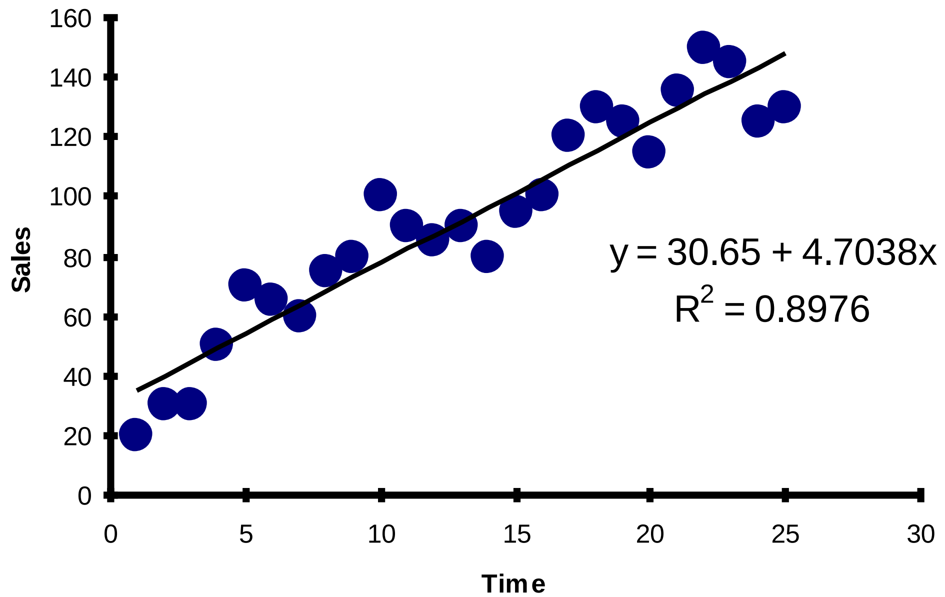
  (for sample size **n** and number of independent variables **k.**)

Decision rule: reject H_0 if $D < d_L$



Testing for Positive Autocorrelation (continued)

- Suppose we have the following time series data:



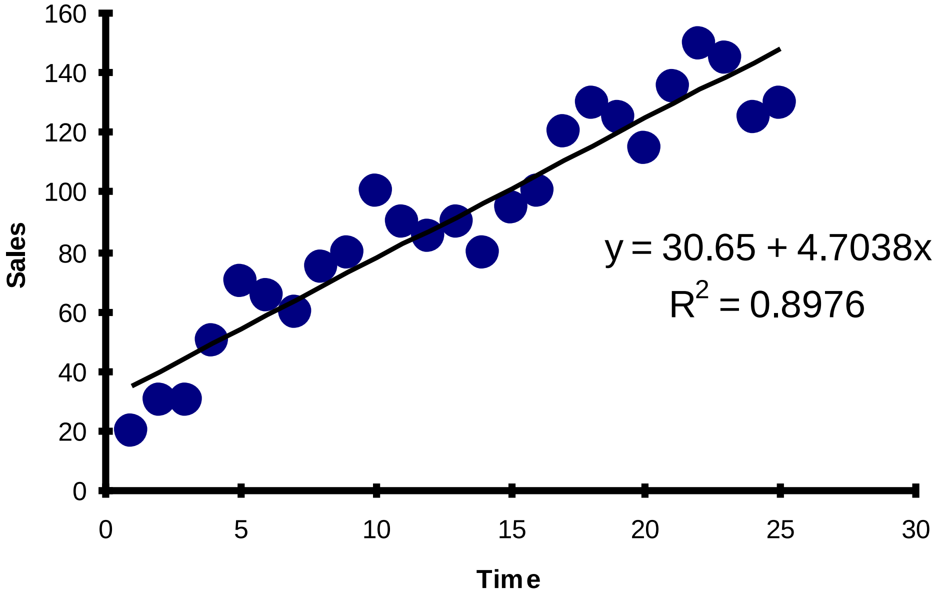
- Is there autocorrelation?

Testing for Positive Autocorrelation
(continued)

- Example with n = 25:

Excel/PHStat output:

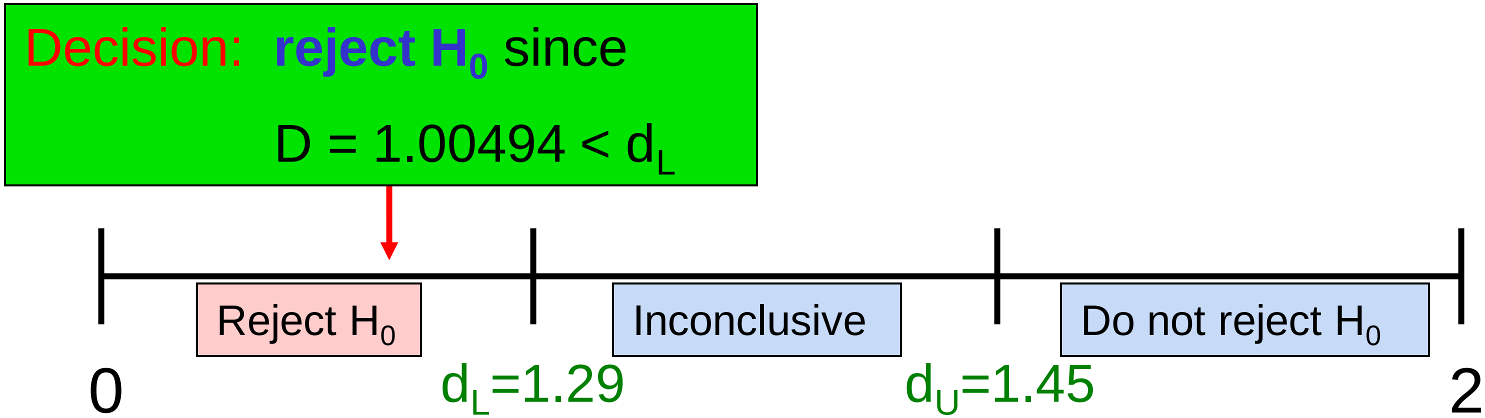
Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**



$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

Testing for Positive Autocorrelation (continued)

- Here, $n = 25$ and there is $k = 1$ one independent variable
- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$
- $D = 1.00494 < d_L = 1.29$, so reject H_0 and conclude that <u>significant positive autocorrelation exists</u>

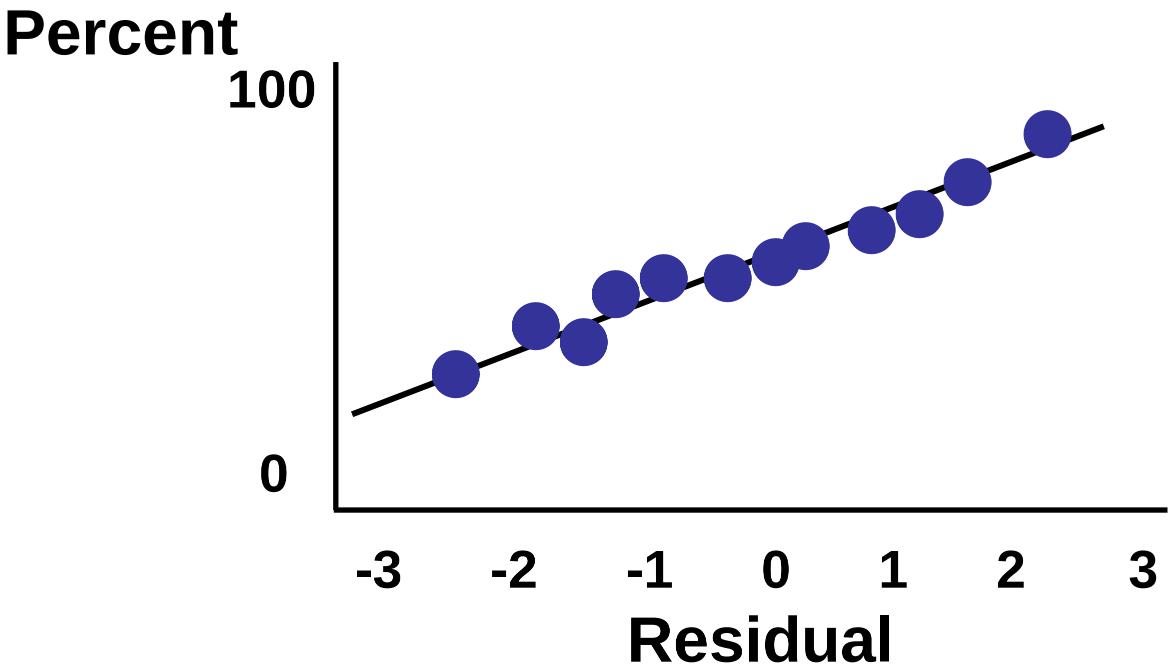


Checking for Normality

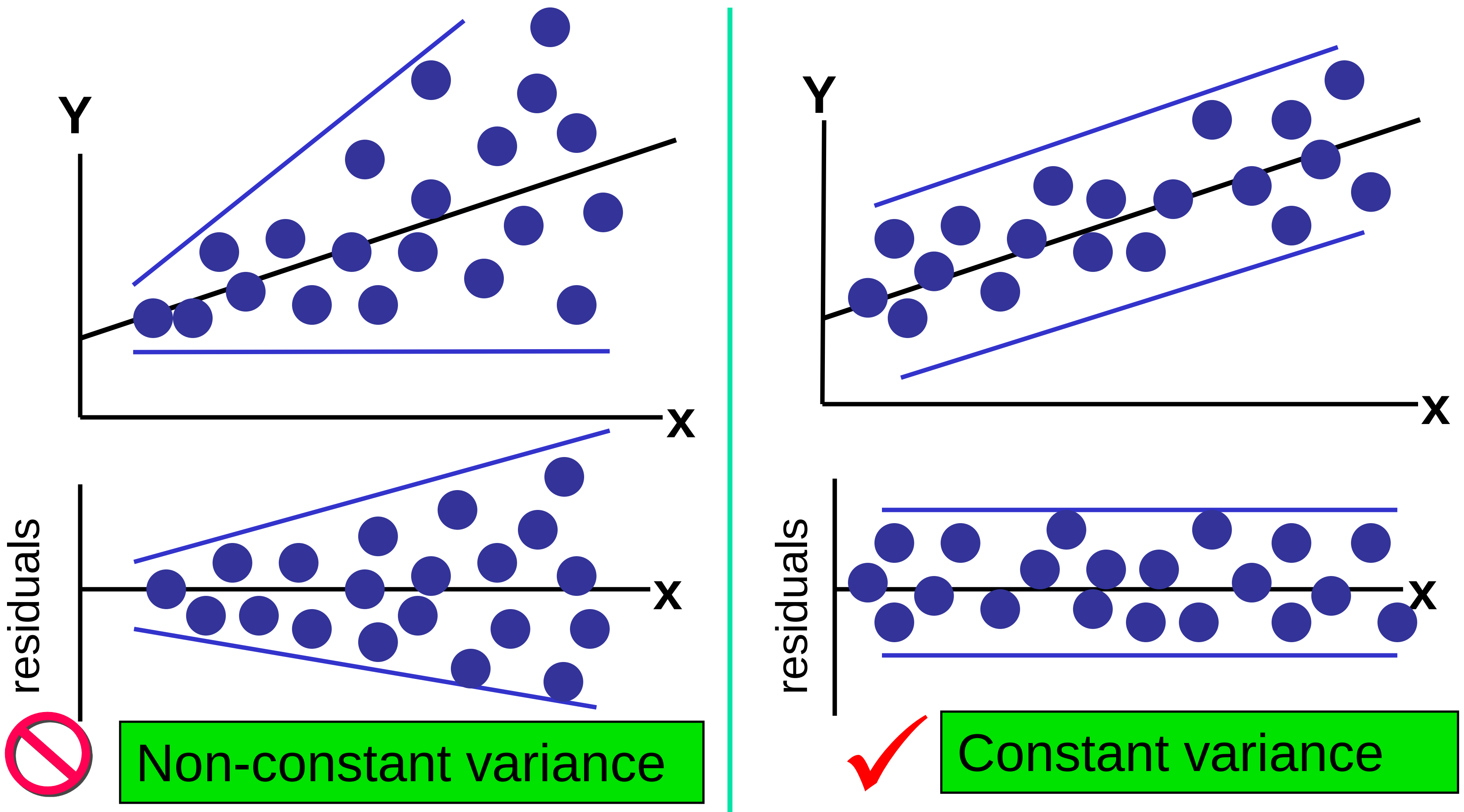
- Examine the Stem-and-Leaf Display of the Residuals.
- Examine the **Boxplot** of the Residuals.
- Examine the Histogram of the Residuals.
- Construct a **Normal Probability Plot** of the Residuals.

Residual Analysis for Normality

When using a normal probability plot, normal errors will approximately display in a straight line.



Residual Analysis for Equal Variance (White test)



Residual Analysis for Equal Variance (White test)

Homoskedasticity: the variance of the error term is constant.
Heteroskedasticity: Absence of Homoscedasticity – or, If the error terms do not have constant variance

Heteroskedasticity test: White test

The null hypothesis for White's test is that the variances for the errors are equal. In math terms, that's:
$H_0 : \sigma^2_i = \sigma^2$.
The alternate hypothesis (the one you're testing), is that the variances are not equal:
$H_1 : \sigma^2_i \neq \sigma^2$.

Residual Analysis for Equal Variance (White test)

1 Estimate your model using OLS:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

2 Obtain the predicted *Y* values after estimating your model.

3 Estimate the model using OLS:

$$\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i + \delta_2 \hat{Y}_i^2$$

4 Retain the R-squared value from this regression:

$$R^2_{\hat{\varepsilon}^2}$$

5 Calculate the F-statistic or the chi-squared statistic:

$$F = \frac{\frac{R^2_{\hat{\varepsilon}^2}}{1}}{\frac{\left(1 - R^2_{\hat{\varepsilon}^2}\right)}{n-2}} \text{ or } \chi^2 = nR^2_{\hat{\varepsilon}^2}$$

The degrees of freedom for the F-test are equal to 2 in the numerator and n – 3 in the denominator. The degrees of freedom for the chi-squared test are 2.

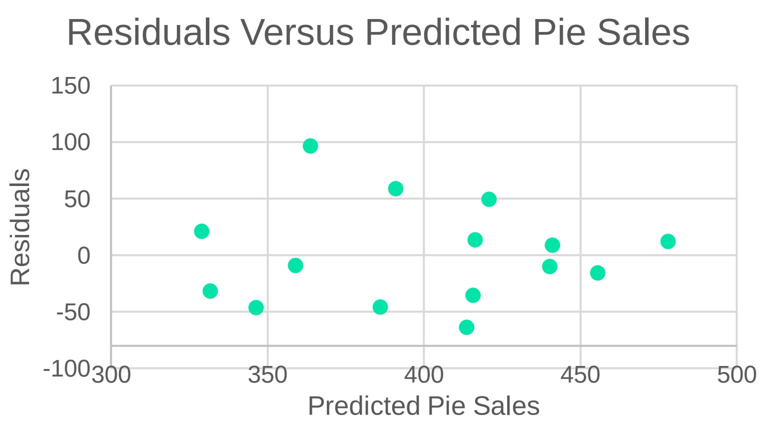
If either of these test statistics is significant, then you have evidence of heteroskedasticity. If not, you fail to reject the null hypothesis of homoskedasticity.

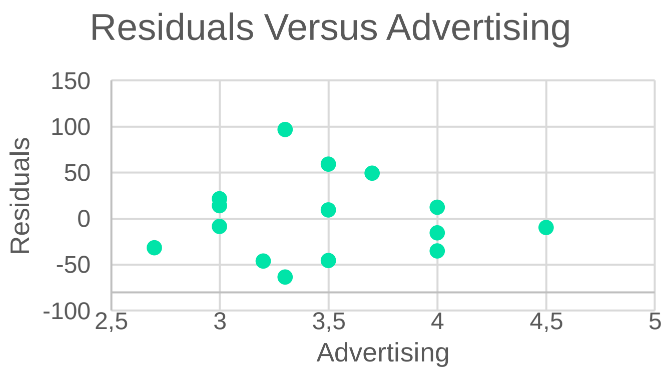
Residual Plots Used in Multiple Regression

- 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 – autocorrelation check).

Use the residual plots to check for violations of regression assumptions.

Residual Plots For The Pie Sales Model





All these plots show little or no pattern so we can conclude the multiple regression model is appropriate for predicting pie sales.

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)

Are Individual Variables Significant?

(continued)

H_0: $\beta_j = 0$ (no linear relationship between X_j and Y)

H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Test Statistic:

$$\boxed{t_{STAT} = \frac{b_j - 0}{S_{b_j}}}$$

(df = n – k – 1)

Inferences about the Slope: t Test Example

H_0: $\beta_j = 0$
H_1: $\beta_j \neq 0$

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

α = .05

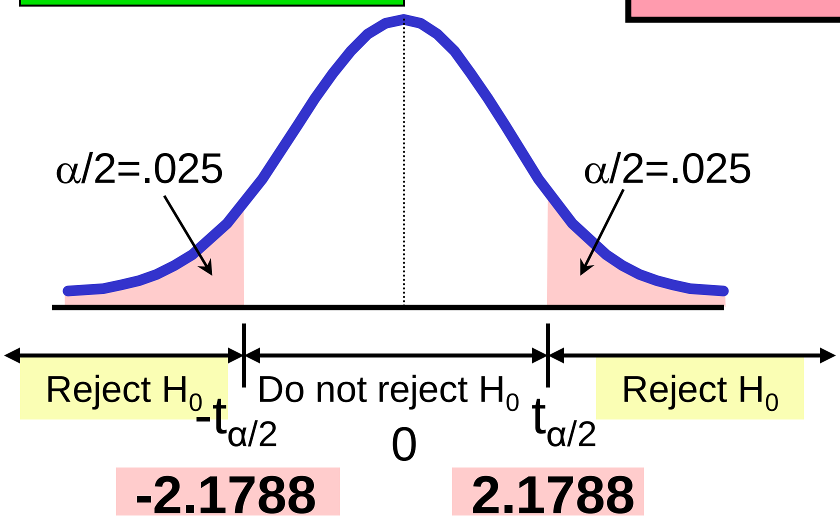
$t_{\alpha/2}$ = 2.1788

From the Excel output:

For Price t_{STAT} = -2.306, with p-value .0398.

For Advertising t_{STAT} = 2.855, with p-value .0145

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



Decision:

Reject H_0 for each variable.

Conclusion:

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

Confidence Interval Estimate for the Slope

Confidence interval for the population slope β_j

$$\boxed{b_j \pm t_{\alpha/2} S_{b_j}}$$

where t has (n – k – 1) d.f.

Here, t has

(15 – 2 – 1) = 12 d.f.

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).

Testing Portions of the Multiple Regression Model

- Contribution of a Single Independent Variable X_j.

$$SSR(X_j \mid \text{all variables except } X_j)$$
$$= SSR\,(\text{all variables}) - SSR(\text{all variables except } X_j)$$

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

Testing Portions of the Multiple Regression Model

(continued)

Contribution of a Single Independent Variable X_j, assuming all other variables are already included (consider here a 2-variable model):

$SSR(X_1 \mid X_2)$

$= SSR \text{ (all variables)} - SSR(X_2)$

From ANOVA section of regression for

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

From ANOVA section of regression for

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

Measures the contribution of X_1 in explaining SST.

The Partial F-Test Statistic

- Consider the hypothesis test:

H_0: variable X_j does not significantly improve the model after all other variables are included

H_1: variable X_j significantly improves the model after all other variables are included

- Test using the F-test statistic:

(with 1 and n-k-1 d.f.)

$$F_{STAT} = \frac{\text{SSR}\ (X_j \mid \text{all variables except j})}{\text{MSE}}$$

Testing Portions of Model: Example

Example: Frozen dessert pies

Test at the $\alpha = .05$ level to determine whether the price variable significantly improves the model given that advertising is included.

Testing Portions of Model: Example

(continued)

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

H_1: X_1 does improve model

$\alpha = .05$, df = 1 and 12

$F_{0.05} = 4.75$

(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

Testing Portions of Model: Example

(continued)

(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

$$F_{STAT} = \frac{SSR\,(X_1 \mid X_2)}{MSE(\text{all})} = \frac{29{,}460.03 - 17{,}484.22}{2{,}252.78} = 5.316$$

Conclusion: Since $F_{STAT} = 5.316 > F_{0.05} = 4.75$ Reject H_0;
Adding X_1 does improve model.

Testing Portions of Model: Example

(continued)

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

H_1: X_2 does improve model

$\alpha = .05$, df = 1 and 12

$F_{0.05} = 4.75$

(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_1 only)

ANOVA		
	df	*SS*
Regression	1	**11100.43803**
Residual	13	45392.8953
Total	14	56493.33333

Testing Portions of Model: Example

(continued)

(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_1 only)

ANOVA		
	df	*SS*
Regression	1	**11100.43803**
Residual	13	45392.8953
Total	14	56493.33333

$$F_{STAT} = \frac{SSR\,(X_2 \mid X_1)}{MSE(all)} = \frac{29{,}460.03 - 11{,}100.44}{2{,}252.78} = 8.150$$

Conclusion: Since $F_{STAT} = 8.150 > F_{0.05} = 4.75$ Reject H_0;
Adding X_2 does improve model.

Relationship Between Test Statistics

- The partial *F* test statistic developed in this section and the *t* test statistic are both used to determine the contribution of an independent variable to a multiple regression model.
- The hypothesis tests associated with these two statistics always result in the same decision (that is, the *p*-values are identical).

$$t_{STAT}^2 = F_{STAT}$$

Tip: in practice, you can use directly the p-values from Excel related to each variable to reach the same conclusion

Coefficients of Partial Determination for 2 variable model

$$r^2_{Y1.2} = \frac{SSR(X_1|X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2|X_1)}$$

- Measures the proportion of variation in Y explained by X_1 while controlling for (holding constant) X_2.

$$r^2_{Y2.1} = \frac{SSR(X_2|X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1|X_2)}$$

- Measures the proportion of variation in Y explained by X_2 while controlling for (holding constant) X_1.

Coefficients of Partial Determination for Pie Sales Data Using JMP Output

Response Pie Sales

Effect Summary

Summary of Fit

RSquare	0.521478
RSquare Adj	0.441724
Root Mean Square Error	47.46341
Mean of Response	399.3333
Observations (or Sum Wgts)	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	29460.027	14730.0	6.5386
Error	12	27033.306	2252.8	Prob > F
C. Total	14	56493.333		0.0120*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	306.52619	114.2539	2.68	0.0199*
Price	-24.97509	10.83213	-2.31	0.0398*
Advertising	74.130957	25.96732	2.85	0.0145*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Price	1	1	11975.804	5.3160	0.0398*
Advertising	1	1	18359.589	8.1498	0.0145*

Effect Details

$$r^2_{Y1.2} = \frac{11{,}975.804}{54{,}493.333 - 29{,}460.027 + 18{,}359.589} = 0.276$$

$$r^2_{Y2.1} = \frac{18{,}359.589}{54{,}493.333 - 29{,}460.027 + 11{,}975.804} = 0.496$$

Using Dummy Variables

- A dummy variable is a categorical independent 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 - 1).

Dummy-Variable Example (with 2 Levels)

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

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).

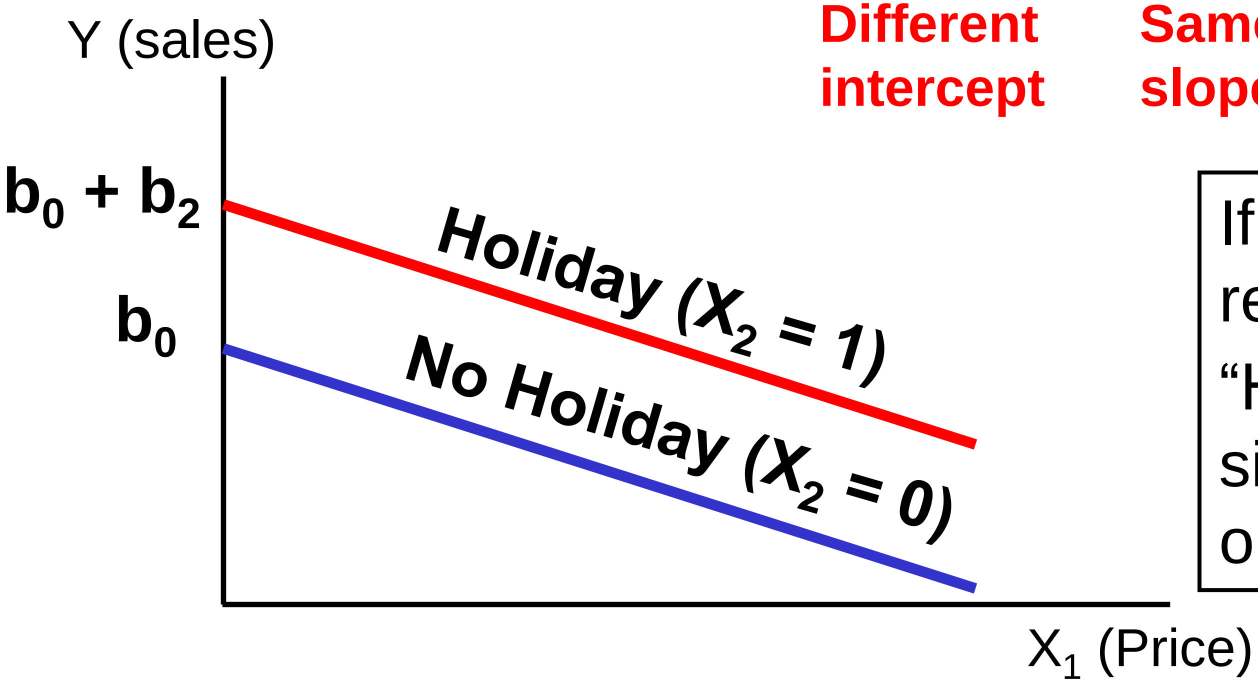
Dummy-Variable Example (with 2 Levels)

(continued)

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

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

Different intercept **Same slope**



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

Dummy-Variable Example Excel Output

(continued)

Regression Statistics	
Multiple R	0.785273784
R Square	0.616654916
Adjusted R Squ	0.552764069
Standard Error	42.4818016
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	34836.89173	17418.45	9.651694	0.003173506
Residual	12	21656.4416	1804.703		
Total	14	56493.33333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	548.0984369	65.08044138	8.421861	2.21E-06	406.3003363	689.8965376
Price	-26.12839289	9.707921784	-2.69145	0.019617	-47.2801374	-4.97664836
Holiday	90.11500528	24.84803909	3.626645	0.003472	35.97577893	144.2542316

Interpreting the Dummy Variable Coefficient (with 2 Levels)

$$\widehat{Sales} = 548.1 - 26.1(\text{Price}) + 90.1(\text{Holiday})$$

Sales: number of pies sold per week

Price: pie price in $

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

Interpretation:

b_2 = 90.1: on average, sales were 90.1 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 = house price ; X_1 = square feet.

- If style of the house is also thought to matter:

  Style = $\underbrace{\text{ranch, split level, colonial}}$.

Three levels, so two dummy variables are needed.

Dummy-Variable Models (more than 2 Levels)

(continued)

- Example: Let "colonial" 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:

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

Interpreting the Dummy Variable Coefficients (with 3 Levels)

Consider the regression equation:

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

For a colonial: $X_2 = X_3 = 0$

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

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 colonial.

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 colonial.

Example of multiple regression - Cars

A car’s power output is influenced by several factors. A sample of 90 different car models of three makes from the EU market is stored in the file Cars. Develop a multiple linear regression model to predict power output (kW), based on engine size (displacement in cubic centimeters) and maximum speed (km/h) [14.5 from Berenson Book]

a) State the multiple regression equation
b) Interpret the meaning of the slopes, b1 & b2 in this problem
c) Explain why the regression coefficient b0 has no practical meaning in the context of this problem
d) Predict the mean power output of cars that have a displacement of 1800 cm3 and max speed of 200 km/h

Solution in Excel

Example of multiple regression – Cars Solution

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,915589967
R Square	0,838304988
Adjusted R Square	0,834587861
Standard Error	25,09286232
Observations	90

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	284004,4126	142002,2063	225,5249965	3,78707E-35
Residual	87	54779,70131	629,6517392		
Total	89	338784,1139			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	-205,609595	17,75483417	-11,58048524	2,60148E-19	-240,8992505	-170,3199396	-240,8992505	-170,3199396
Displacement (cm³)	0,033672542	0,006277992	5,363584648	6,64591E-07	0,021194353	0,046150731	0,021194353	0,046150731
Max speed (km/h)	1,230694485	0,116305304	10,58158523	2,65395E-17	0,9995251	1,46186387	0,9995251	1,46186387

(a) $\hat{Y} = -205.6096 + 0.0337 X_1 + 1.2307 X_2$

(b) For a given maximum speed, each unit increase of the displacement is estimated to result in an increase in mean power output of 0.0337 kW. For a given displacement, each unit increase of the maximum speed is estimated to result in an increase in mean power output of 1.230 kW.

(c) The interpretation of b_0 has no practical meaning here because it would represent the power output when there were no displacement and zero maximum speed, which is obviously not possible.

(d) $\hat{Y} = -205.6096 + 0.0337(1800) + 1.2307(200) = 101.1399$ (kw)

How to standardize a variable

- Standardization means having a mean of zero and unitary variance
- To achieve this, we need calculate mean m and standard deviation s

The standardized value of the variable X is equal to

$$\frac{X_i - m}{s}$$

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.
  - $$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$
    $$= b_0 + b_1X_1 + b_2X_2 + b_3(X_1X_2)$$

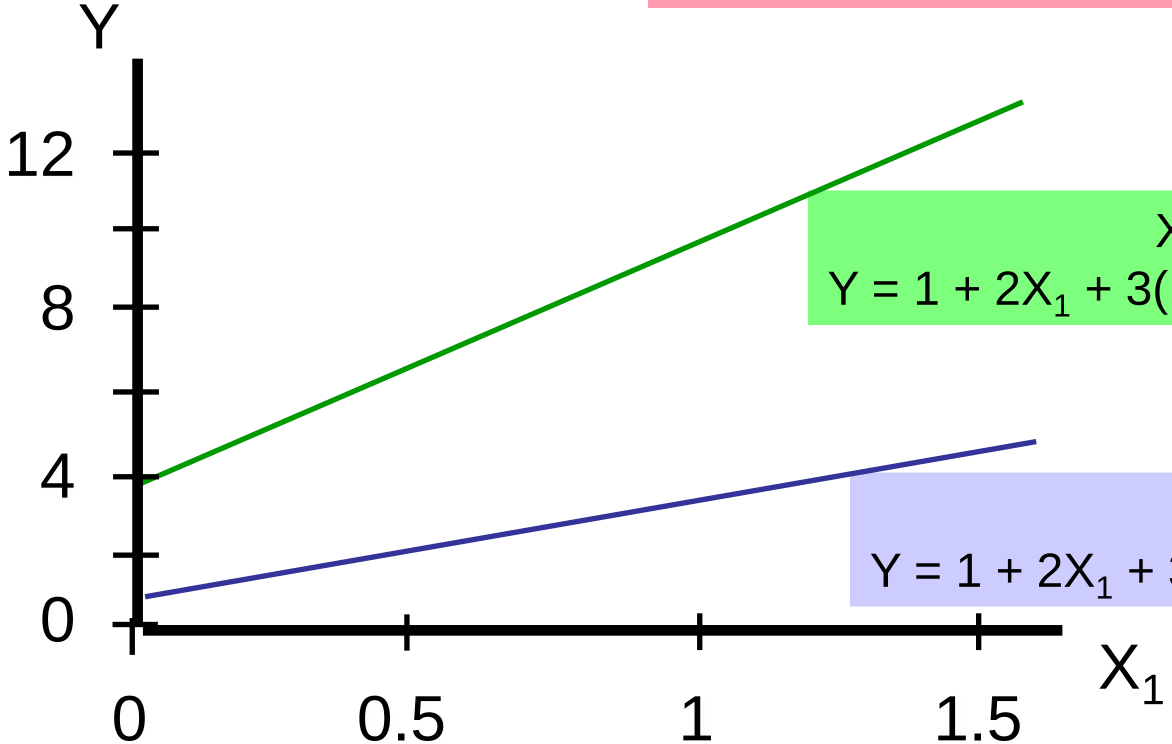
Effect of Interaction

- 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 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.

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

Example of Interaction Term

- Consider example of asking price of homes from pages 589 – 591.
- Dependent variable is Asking Price of Homes.
- Independent variables are Living Space and whether or not the house has a fireplace.
- In the model with no interaction you are assuming the effect of Living Space on Asking Price is the same whether or not the house has a fireplace.
- If this is not true, the model needs an interaction term.

Excel, Output From Interaction Model

16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	316.2350	50.7878	6.2266	0.0000	214.5341	417.9359
18	Living Space	0.0681	0.0292	2.3319	0.0233	0.0096	0.1265
19	Fireplace	34.8926	59.0359	0.5910	0.5568	-83.3248	153.1101
20	Living Space*Fireplace	0.0106	0.0326	0.3239	0.7472	-0.0548	0.0759

Interaction is not significant.

Effect of Living Space on Asking Price does not depend on whether or not the house has a fireplace.

If the interaction was significant, we would interpret it as: Changes in Living space have a significant effect on Asking price only for those houses that have a fireplace

Evaluating A Model With Several Interactions Simultaneously

- In a model with multiple interaction terms, can use a partial F test to evaluate their contribution.
- Consider example 14.4 where:
  - Y = monthly heating oil consumption
  - X_1 = temperature
  - X_2 = insulation
  - X_3 = 1 if house is a ranch, = 0 otherwise
  - $X_4 = X_1 * X_2$
  - $X_5 = X_1 * X_3$
  - $X_6 = X_2 * X_3$
- The regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i$$

Evaluating A Model With Several Interactions Simultaneously (Con't)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i$$

- To decide if any of the interactions are significant we test:

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \ \text{ vs}$$
$$H_1 : \beta_4 \text{ and/or } \beta_5 \text{ and/or } \beta_6 \neq 0$$

- Using a partial F test with m = 3 (the three parameters).

$$F_{STAT} = \frac{[\text{SSR(all)} - \text{SSR (all except new set of m variables)}] / \text{m}}{\text{MSE(all)}}$$

(where F_{STAT} has m and n-k-1 d.f.)

Evaluating A Model With Several Interactions Simultaneouslly (Con't)

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \ \text{ vs}$$

$$H_1 : \beta_4 \text{ and/or } \beta_5 \text{ and/or } \beta_6 \neq 0$$

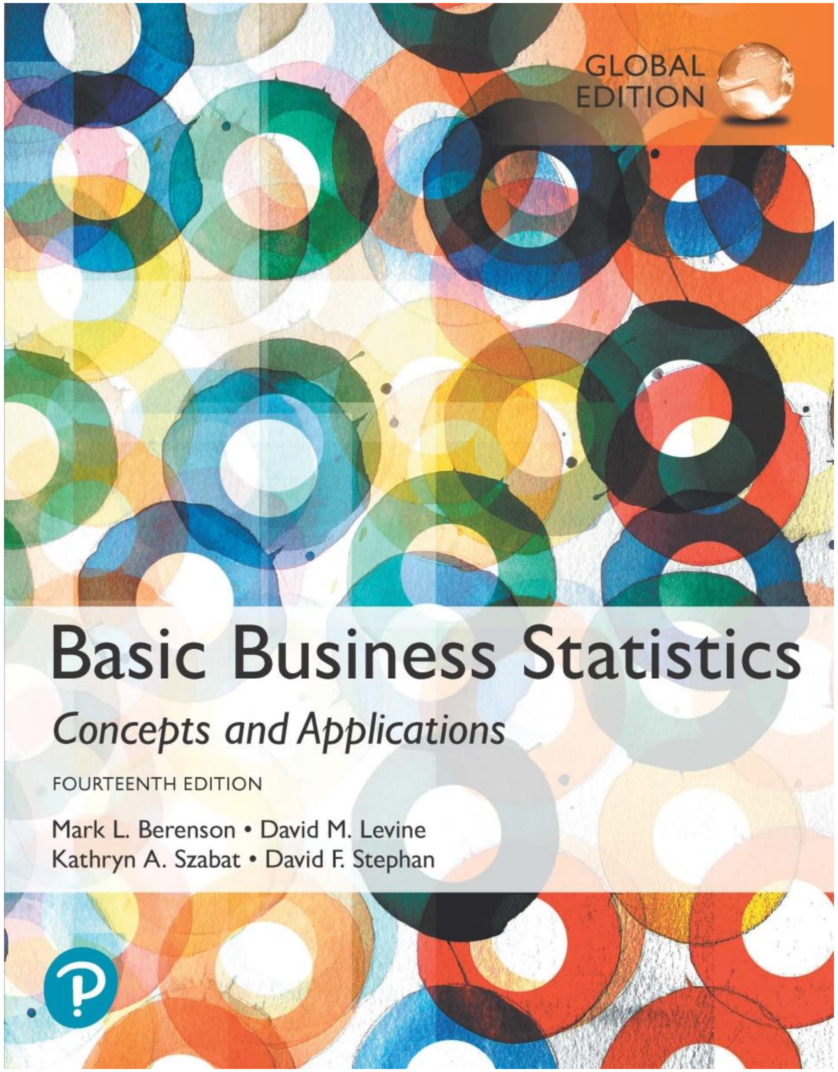
$$F_{STAT} = \frac{[\text{SSR(all)} - \text{SSR (all except new set of m variables)}] / \text{m}}{\text{MSE(all)}}$$

$$= [(234{,}510.5818 - 233{,}406.9094) / 3] / 203.0809 = 1.8115$$

$$\text{Since } F_{STAT} = 1.8115 < F_{0.05,8,3} = 4.07$$

The interactions DO NOT add significant value

If we had rejected H0, then we would need to check individually each interaction to conclude which ones to include in the model



Chapter 15

Multiple Regression Model Building

Nonlinear Relationships

- The relationship between the dependent variable and an independent variable may not be linear.
- Can review the scatter plot to check for non-linear relationships.
- Example: Quadratic model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

  - The second independent variable is the square of the first independent variable.

Quadratic Regression Model

Model form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

- where:

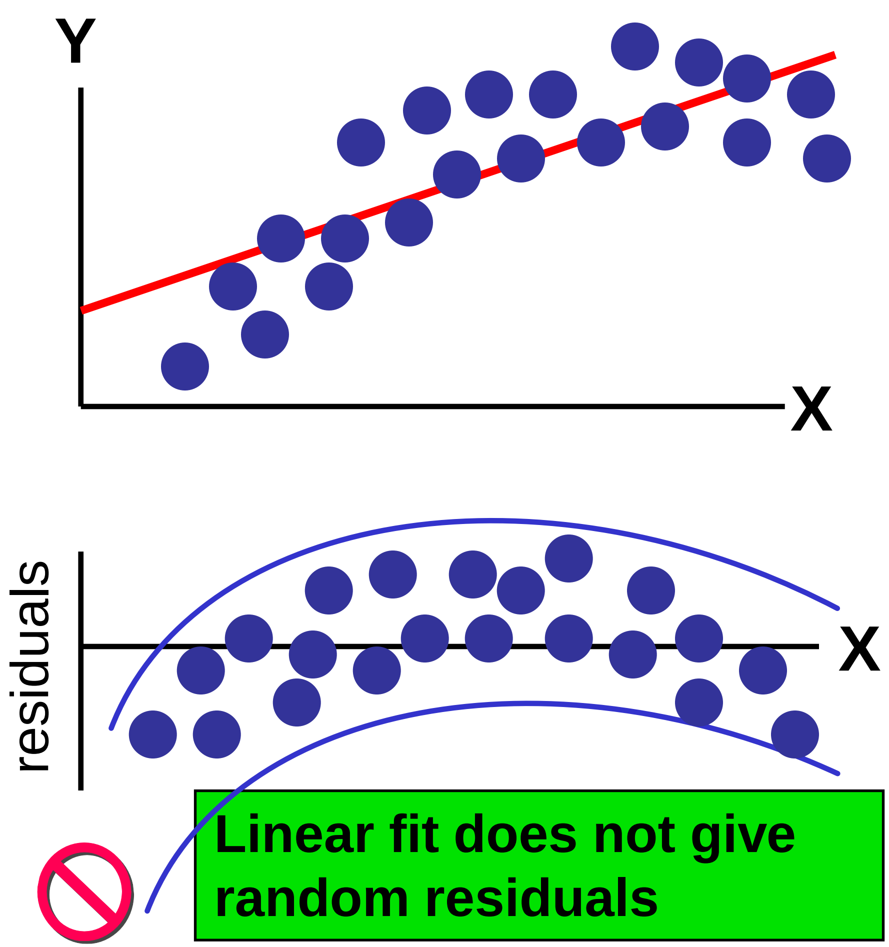
  β_0 = Y intercept

  β_1 = coefficient for linear effect of X on Y

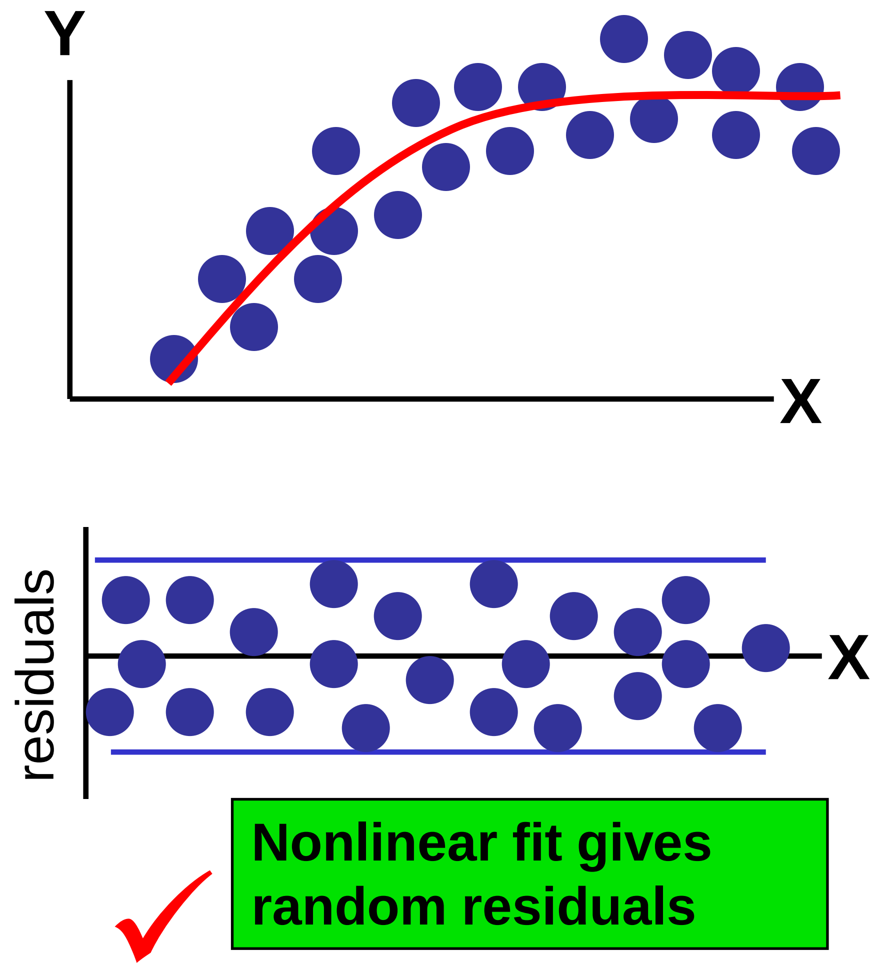
  β_2 = coefficient for quadratic effect on Y

  ε_i = random error in Y for observation i

Linear vs. Nonlinear Fit



Linear fit does not give random residuals

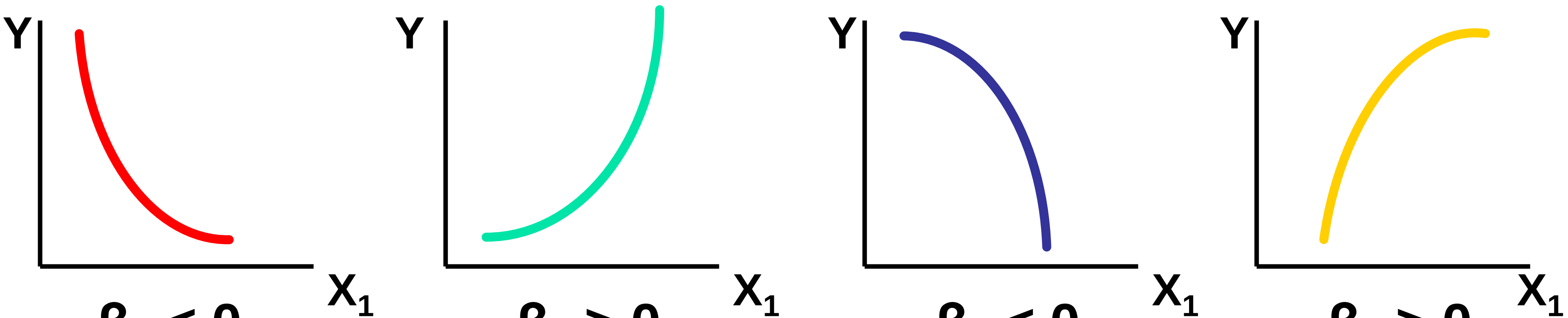


Nonlinear fit gives random residuals

Quadratic Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Quadratic models may be considered when the scatter plot takes on one of the following shapes:



$\beta_1 < 0$	$\beta_1 > 0$	$\beta_1 < 0$	$\beta_1 > 0$
$\beta_2 > 0$	$\beta_2 > 0$	$\beta_2 < 0$	$\beta_2 < 0$

β_1 = the coefficient of the linear term
β_2 = the coefficient of the squared term

Collinearity (you don't need to know the details for this class)

- Collinearity: High correlation exists among two or more independent variables.
- This means the correlated variables contribute redundant information to the multiple regression model.

Model Building (Only the basic info are relevant for this course)

- Goal is to develop a model with the best set of independent variables.
- Model is easier to interpret if unimportant variables are removed.
- Lower probability of collinearity.

Analysis Continues Looking For An Adequate Subset Of Dependent Variables

- Two model building methods are commonly used to study multiple regression models:
  - Stepwise regression procedure (forward [in Berenson et al.] or backward):
    - Provide evaluation of alternative models as variables are added and deleted.
  - Best-subset approach *(just for info)* [in Berenson et al.] :
    - Try all combinations and select the best using the highest adjusted r^2 and C_p criteria.

There is also

- *Forward Selection (just for info)*
- *Backward Selection (just for info)*

Stepwise Regression

- Develop the least squares regression equation in steps.
- Add one independent variable at a time in *forward* stepwise (or start with all variables and remove one at a time in *backward*) and evaluate whether existing variables should remain in the model or be removed.
- The coefficient of partial determination is the measure of the marginal contribution of each independent variable, given that other independent variables are in the model.
- You can also judge by the p-value

Stepwise Regression – Step 1

- Always remember that the Intercept should NEVER be ELIMINATED (no matter what the p-value is)

Regression Statistics	
Multiple R	0,75691347
R Square	0,57291799
Adjusted R Square	0,53097244
Standard Error	8043,05943
Observations	124

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	9719451729	883586520,8	13,6586107	3,0706E-16
Residual	112	7245370160	64690805		
Total	123	16964821889			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	3887,29636	9072,68714	0,428461414	0,6691381	-14089,0702	21863,66294	-14089,0702	21863,66294
STARS	857,915461	894,2065255	0,959415344	0,33941614	-913,840145	2629,671068	-913,840145	2629,671068
Total_Rooms	-70,2091536	18,22527611	-3,852295742	0,00019542	-106,320202	-34,09810516	-106,320202	-34,09810516
ID_Crete	767,792116	1939,406822	0,39589018	0,69293919	-3074,8939	4610,47813	-3074,8939	4610,47813
ID_Southern_Aegean	-103,013369	1897,329206	-0,054293882	0,95679775	-3862,32798	3656,301243	-3862,32798	3656,301243
ARR_MAY	150,896022	55,60814271	2,713559833	0,00770895	40,7156158	261,0764291	40,71561581	261,0764291
ARR_AUG	-4,11943802	38,23163387	-0,107749463	0,91438722	-79,8705198	71,63164376	-79,8705198	71,63164376
OR_MAY	98,523161	35,79526394	2,752407727	0,00690292	27,599434	169,4468881	27,59943398	169,4468881
OR_AUG	-76,8787865	101,4297088	-0,757951367	0,45007197	-277,848753	124,0911797	-277,848753	124,0911797
Total_Empl_May	122,86642	83,94461965	1,463660455	0,14608772	-43,4590832	289,1919236	-43,4590832	289,1919236
Total_Empl_Aug	-96,0243837	93,94677913	-1,022114697	0,3089293	-282,167884	90,11911655	-282,167884	90,11911655
L_COST	0,01092353	0,003948695	2,766365343	0,00663259	0,0030997	0,018747368	0,0030997	0,018747368

Stepwise Regression – Step 2

- Always remember that the Intercept is NEVER ELIMINATED (no matter what the p-value is)

Regression Statistics								
Multiple R	0,75690604							
R Square	0,57290675							
Adjusted R Square	0,53511089							
Standard Error	8007,49696							
Observations	124							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	10	9719261032	971926103,2	15,1579225	7,6304E-17			
Residual	113	7245560857	64120007,58					
Total	123	16964821889						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3821,03131	8950,469898	0,426908459	0,67025829	-13911,4639	21553,52648	-13911,4639	21553,52648
STARS	859,773432	889,6006394	0,966471239	0,33587234	-902,685882	2622,232747	-902,685882	2622,232747
Total_Rooms	-70,4044186	17,78790553	-3,957993731	0,00013244	-105,645468	-35,1633695	-105,645468	-35,1633695
ID_Crete	811,547889	1756,233792	0,462095589	0,64490123	-2667,86801	4290,963784	-2667,86801	4290,963784
ARR_MAY	151,066352	55,27409834	2,733040553	0,00728532	41,5583943	260,5743103	41,55839427	260,5743103
ARR_AUG	-4,50361159	37,40501994	-0,120401262	0,90437911	-78,6097024	69,60247925	-78,6097024	69,60247925
OR_MAY	98,6526316	35,55782806	2,774427938	0,00647214	28,206161	169,0991022	28,20616101	169,0991022
OR_AUG	-76,5034349	100,7464026	-0,759366418	0,44921461	-276,100229	123,0933593	-276,100229	123,0933593
Total_Empl_May	122,549989	83,37179174	1,469921503	0,14436186	-42,6245669	287,7245457	-42,6245669	287,7245457
Total_Empl_Aug	-95,4155562	92,86273217	-1,027490296	0,30638406	-279,393377	88,5622649	-279,393377	88,5622649
L_COST	0,01093487	0,003925736	2,785431651	0,00627033	0,00315728	0,018712462	0,003157279	0,018712462

Pearson ALWAYS LEARNING

Stepwise Regression – Step 3

- Always remember that the Intercept is NEVER ELIMINATED (no matter what the p-value is)

Regression Statistics								
Multiple R	0,75686985							
R Square	0,57285196							
Adjusted R Square	0,53912975							
Standard Error	7972,81036							
Observations	124							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	9	9718331518	1079814613	16,9873773	1,7964E-17			
Residual	114	7246490370	63565705					
Total	123	16964821889						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3917,05466	8876,250286	0,441296102	0,65983449	-13666,7288	21500,83811	-13666,7288	21500,83811
STARS	851,925181	883,3661088	0,964407818	0,33688285	-898,016276	2601,866639	-898,016276	2601,866639
Total_Rooms	-70,6838503	17,55945821	-4,025400412	0,00010276	-105,469002	-35,89869885	-105,469002	-35,89869885
ID_Crete	839,045868	1733,777455	0,483940927	0,62935614	-2595,55392	4273,645654	-2595,55392	4273,645654
ARR_MAY	145,357993	28,29189904	5,137795563	1,1607E-06	89,3119593	201,4040274	89,31195925	201,4040274
OR_MAY	98,1681053	35,17633229	2,790743062	0,00616665	28,4840619	167,8521486	28,48406194	167,8521486
OR_AUG	-77,9123509	99,63104638	-0,782008758	0,43583065	-275,280686	119,4559847	-275,280686	119,4559847
Total_Empl_May	122,396005	83,0008772	1,474635081	0,14306702	-42,0280932	286,8201038	-42,0280932	286,8201038
Total_Empl_Aug	-95,7388382	92,42181493	-1,035890047	0,3024458	-278,825742	87,34806534	-278,825742	87,34806534
L_COST	0,01093471	0,003908731	2,797510043	0,00604719	0,00319155	0,01867788	0,003191548	0,01867788

Stepwise Regression – Step … n (Final)

- Always remember that the Intercept is NEVER ELIMINATED (no matter what the p-value is)

Regression Statistics								
Multiple R	0,74424888							
R Square	0,55390639							
Adjusted R Square	0,53891165							
Standard Error	7974,69664							
Observations	124							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	9396923302	2349230825	36,9400389	4,6966E-20			
Residual	119	7567898587	63595786,44					
Total	123	16964821889						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1974,59815	2094,461446	-0,942771306	0,34770796	-6121,841	2172,644698	-6121,841	2172,644698
Total_Rooms	-64,6222438	14,91019892	-4,33409669	3,0777E-05	-94,1459268	-35,09856074	-94,1459268	-35,09856074
ARR_MAY	155,686656	24,90448431	6,251350305	6,5821E-09	106,373289	205,0000219	106,3732893	205,0000219
OR_MAY	105,748821	30,44485806	3,473454218	0,00071728	45,464961	166,0326803	45,46496096	166,0326803
L_COST	0,01125859	0,00225257	4,998110904	2,0127E-06	0,00679828	0,015718907	0,006798281	0,015718907

Revisit the example of Cars

You wish to build a multiple regression model to estimate the **fuel consumption** (l/100km) considering the following potential independent variables

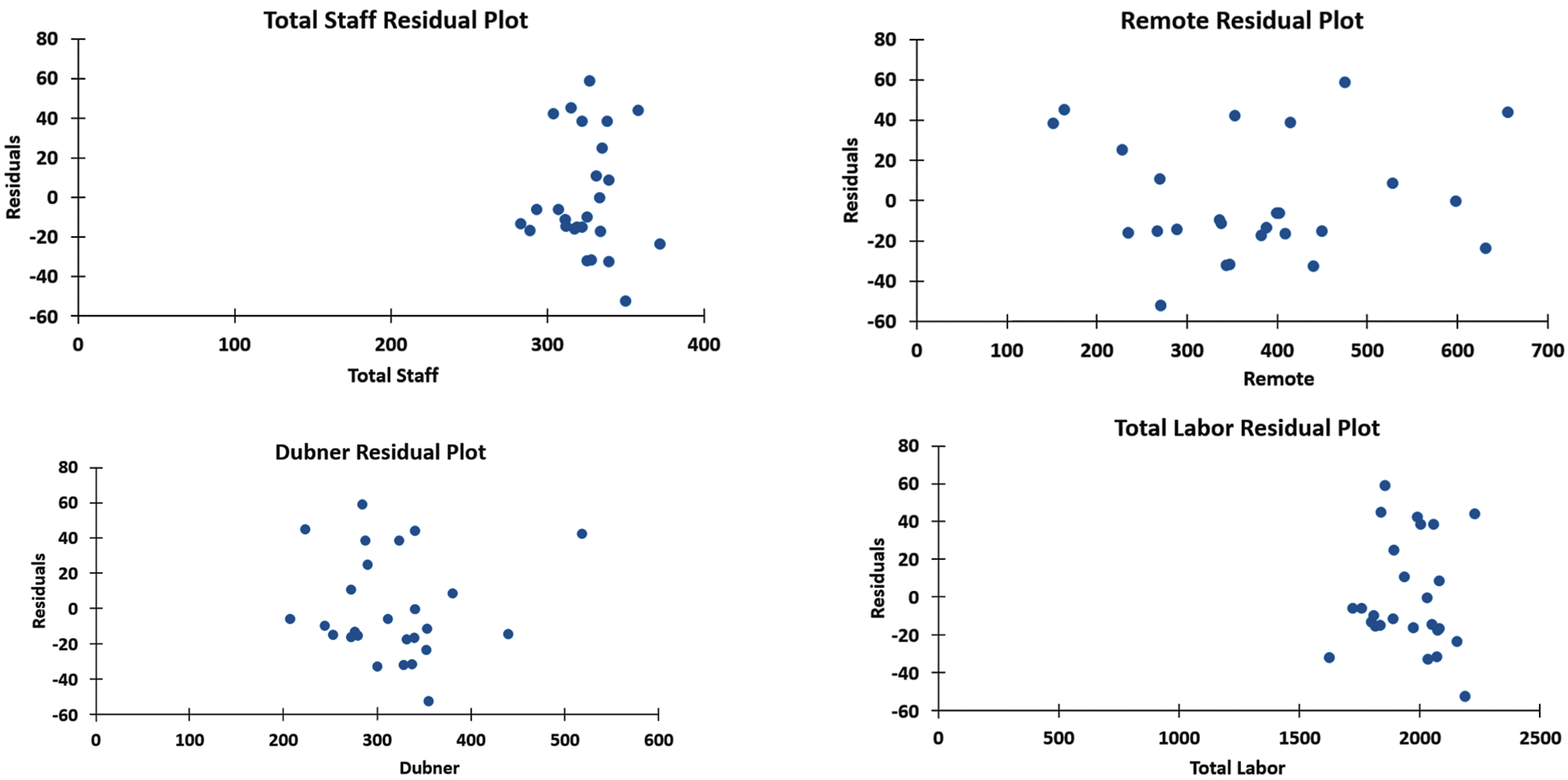
a) Engine displacement (cm3)
b) Max Speed
c) Power (kW)
d) 0-100 kmph (sec)
e) Whether the car is a BMW or not (dummy)

1. Eliminate the non-significant variables,
2. state the multiple regression equation and
3. interpret the meaning of slopes.

Solution in Excel

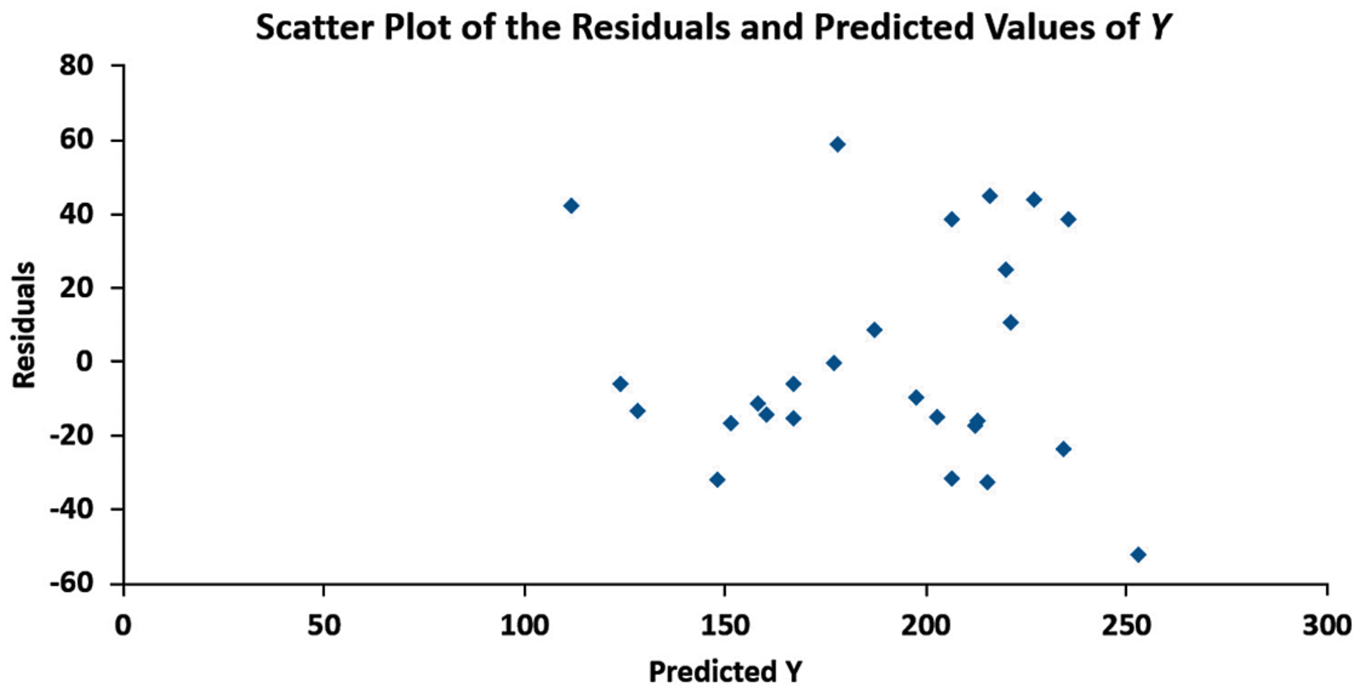
Residual Analysis Should Be Done On The Chosen Model (4 X's)

Residual plots versus each independent variable show only random scatter (no pattern).



Residual Analysis (continued)

Residual versus Predicted Y show constant variance.



The Final Model Building Step Is Model Validation

- Models can be validated via multiple methods:
  - Collect new data and compare the results.
  - Compare the results of the regression model to previous results.
  - If the data set is large enough, split the data into two parts and cross-validate the results.
    - To do this you split the data prior to building the model and use one half of the data to build the model and the other half to validate the model.

Relevant sections from Berenson et al. textbook

- 13.6
- Whole Chapter 14 (except 14.7 & 14.8)
- 15.1 (Quadratic regression) & 15.2 (the Log transformation)
- 15.4 (an overview of methods to build multiple regression models)