Statistics Foundation

Multiple Linear Regression

Digital Vidya

Module Goals

After completing this module, you should be able to:

- Apply multiple regression analysis to business decision-making situations
- Analyze and interpret the computer output for a multiple regression model
- Perform a hypothesis test for all regression coefficients or for a subset of coefficients
- Fit and interpret nonlinear regression models
- Incorporate qualitative variables into the regression model by using dummy variables
- Discuss model specification and analyze residuals

The Multiple Regression Model

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

Multiple Regression Model with k Independent Variables:

Y-intercept

Population slopes

Random Error

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

Multiple Regression Equation

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

Multiple regression equation with k independent variables:

Estimated (or predicted) value of y

Estimated intercept

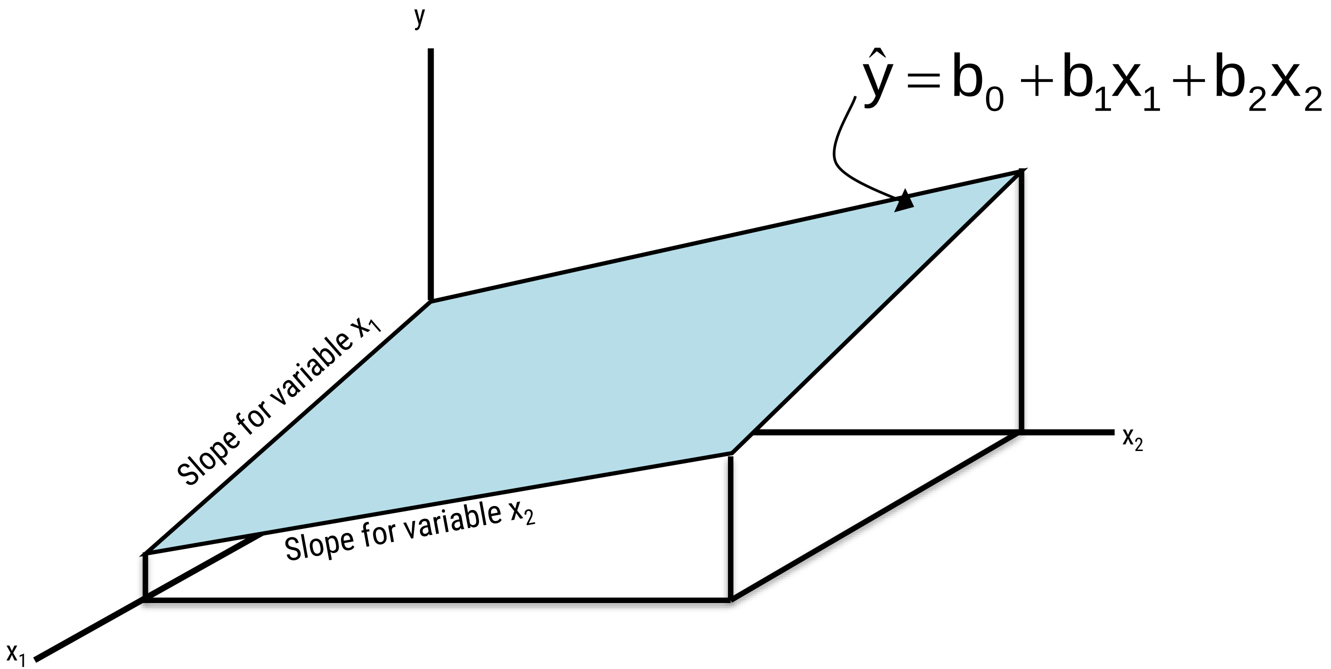
Estimated slope coefficients

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

Multiple Regression Equation

(continued)

Two variable model



Digital Vidya

Standard Multiple Regression Assumptions

- The values x_i and the error terms ε_i are independent

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

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

(The constant variance property is called homoscedasticity)

Standard Multiple Regression Assumptions

(continued)

- The random error terms, ε_i , are not correlated with one another, so that

$$\boxed{E[\varepsilon_i \varepsilon_j] = 0 \quad \text{for all } \ i \neq j}$$

- It is not possible to find a set of numbers, c_0, c_1, . . . , c_k, such that

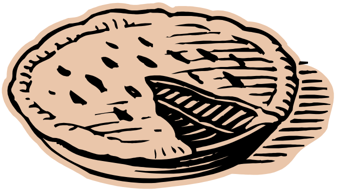
$$\boxed{c_0 + c_1 x_{1i} + c_2 x_{2i} + K + c_K x_{Ki} = 0}$$

(This is the property of no linear relation for the X_j's)

Example:
2 Independent Variables

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

  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in $)
    Advertising ($100's)



- Data is collected for 15 weeks

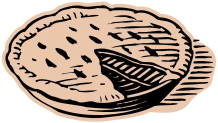
Digital Vidya

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:

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



Digital Vidya

Estimating a Multiple Linear Regression Equation

- Excel will be used to generate the coefficients and measures of goodness of fit for multiple regression

- Excel:
  - Tools / Data Analysis... / Regression
- PHStat:
  - PHStat / Regression / Multiple Regression…

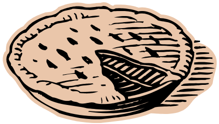
Multiple Regression Output

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

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

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

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

Coefficient of Determination, R^2

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

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

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

Coefficient of Determination, R^2

(continued)

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**	14730.013	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

Estimation of Error Variance

- Consider the population regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \Lambda + \beta_K x_{Ki} + \varepsilon_i$$

- The unbiased estimate of the variance of the errors is

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-K-1} = \frac{SSE}{n-K-1}$$

where $e_i = y_i - \hat{y}_i$

- The square root of the variance, s_e , is called the standard error of the estimate

Digital Vidya

Standard Error, s_e

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

$$s_e = 47.463$$

The magnitude of this value can be compared to the average y value

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

Digital Vidya

Adjusted Coefficient of Determination, $\bar{R}^2$

- R^2 never decreases when a new X variable is added to the model, even if the new variable is not an important predictor variable
  - 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 Coefficient of Determination, $\bar{R}^2$

(continued)

- Used to correct for the fact that adding non-relevant independent variables will still reduce the error sum of squares

$$\bar{R}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}$$

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

  - Adjusted R^2 provides a better comparison between multiple regression models with different numbers of independent variables
  - Penalize excessive use of unimportant independent variables
  - Smaller than R^2

Adjusted Coefficient of Determination, $\bar{R}^2$

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

$$\bar{R}^2 = .44172$$

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	14730.013	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

Digital Vidya

Coefficient of Multiple Correlation

- The coefficient of multiple correlation is the correlation between the predicted value and the observed value of the dependent variable

$$\boxed{R = r(\hat{y}, y) = \sqrt{R^2}}$$

- Is the square root of the multiple coefficient of determination
- Used as another measure of the strength of the linear relationship between the dependent variable and the independent variables
- Comparable to the correlation between Y and X in simple regression

Digital Vidya

Evaluating Individual Regression Coefficients

- Use t-tests for individual coefficients
- Shows if a specific independent variable is conditionally important
- Hypotheses:
  - H_0: $\beta_j = 0$ (no linear relationship)
  - H_1: $\beta_j \neq 0$ (linear relationship does exist between x_j and y)

Evaluating Individual Regression Coefficients

(continued)

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

H_1: $\beta_j \neq 0$ (linear relationship does exist between x_i and y)

Test Statistic:

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

where, (df = n – k – 1)

Evaluating Individual Regression Coefficients

(continued)

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

t-value for Price is t = -2.306, with p-value .0398

t-value for Advertising is t = 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

Example: Evaluating Individual Regression Coefficients

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

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

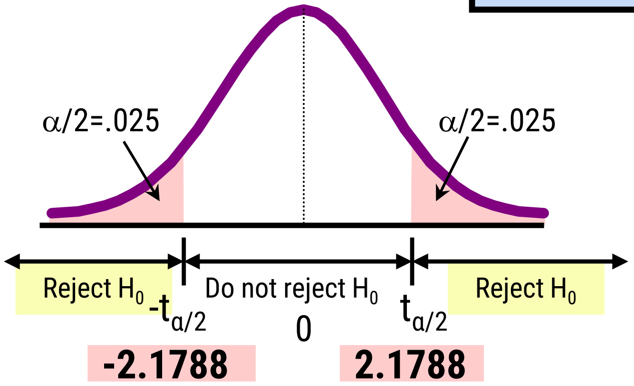
α = .05

$t_{12,\,.025}$ = 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

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 limits for the population slope β_j

$$b_j \pm t_{n-K-1,\alpha/2} S_{b_j}$$

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

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:

$$-24.975 \pm (2.1788)(10.832)$$

So the interval is $-48.576 < \beta_1 < -1.374$

Confidence Interval Estimate for the Slope

(continued)

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:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price

Test on All Coefficients

- 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: $\beta_1 = \beta_2 = \dots = \beta_k = 0$ (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:

$$\boxed{F = \frac{MSR}{s_e^2} = \frac{SSR/K}{SSE/(n-K-1)}}$$

where F has k (numerator) and

(n – K – 1) (denominator)

degrees of freedom

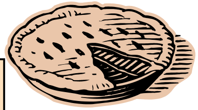
- The decision rule is

$$\boxed{\text{Reject } H_0 \text{ if } F > F_{k,n-K-1,\alpha}}$$

F-Test for Overall Significance

(continued)

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	14730.013	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

F-Test for Overall Significance

(continued)

$H_0: \beta_1 = \beta_2 = 0$

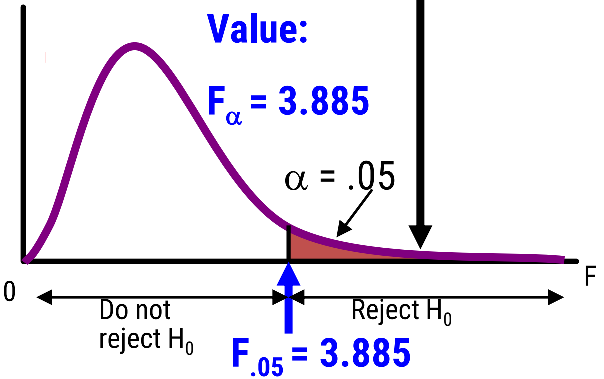
$H_1: \beta_1$ and β_2 not both zero

$\alpha = .05$

$\underbrace{df_1 = 2 \quad df_2 = 12}$

Critical Value:

$F_\alpha = 3.885$



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

Tests on a Subset of Regression Coefficients

- Consider a multiple regression model involving variables x_j and z_j, and the null hypothesis that the z variable coefficients are all zero:

$$\boxed{y_i = \beta_0 + \beta_1 x_{1i} + \Lambda + \beta_K x_{Ki} + \alpha_1 z_{1i} + \Lambda\ \alpha_r z_{ri} + \varepsilon_i}$$

$$\boxed{\begin{array}{l} H_0 : \alpha_1 = \alpha_2 = \Lambda = \alpha_r = 0 \\ H_1 : \text{ at least one of } \alpha_j \neq 0 \quad (j = 1, \ldots, r) \end{array}}$$

Tests on a Subset of Regression Coefficients

(continued)

- Goal: compare the error sum of squares for the complete model with the error sum of squares for the restricted model
  - First run a regression for the complete model and obtain SSE
  - Next run a restricted regression that excludes the z variables (the number of variables excluded is r) and obtain the restricted error sum of squares $SSE(r)$
  - Compute the F statistic and apply the decision rule for a significance level α

$$\text{Reject } H_0 \text{ if } \quad F = \frac{(SSE(r) - SSE)/r}{s_e^2} > F_{r,n-K-r-1,\alpha}$$

Digital Vidya

Prediction

- Given a population regression model

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

- then given a new observation of a data point

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

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

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

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

Using The Equation to Make Predictions

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

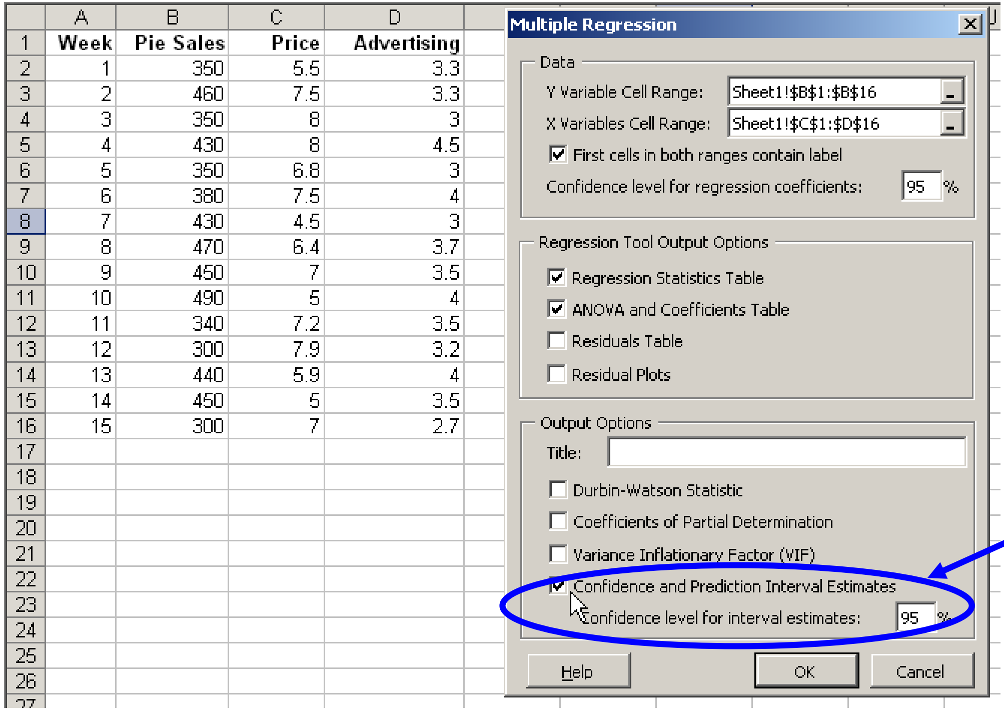
$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$
$$= 306.526 - 24.975\,(5.50) + 74.131\,(3.5)$$
$$= 428.62$$

Predicted sales is 428.62 pies

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

Predictions in PHStat

PHStat | regression | multiple regression …



	A	B	C	D
1	Week	Pie Sales	Price	Advertising
2	1	350	5.5	3.3
3	2	460	7.5	3.3
4	3	350	8	3
5	4	430	8	4.5
6	5	350	6.8	3
7	6	380	7.5	4
8	7	430	4.5	3
9	8	470	6.4	3.7
10	9	450	7	3.5
11	10	490	5	4
12	11	340	7.2	3.5
13	12	300	7.9	3.2
14	13	440	5.9	4
15	14	450	5	3.5
16	15	300	7	2.7

Check the “confidence and prediction interval estimates” box

Predictions in PHStat

(continued)

	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

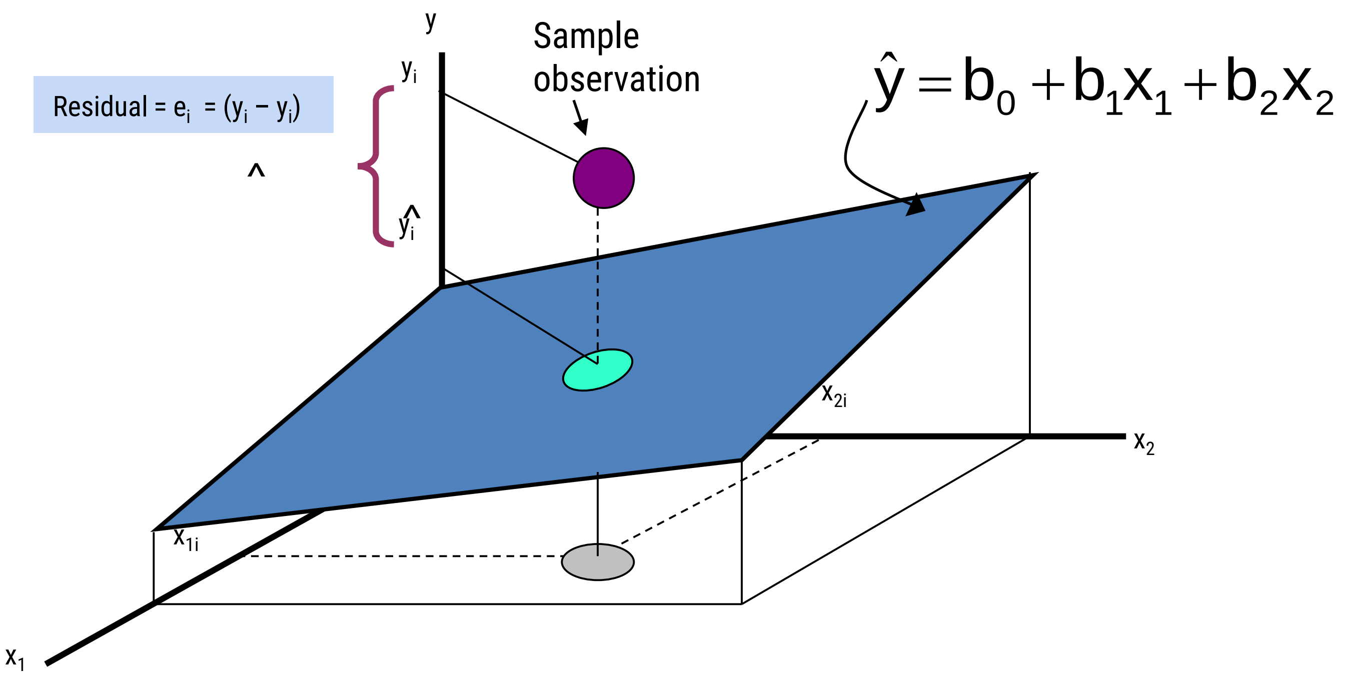
Predicted $\hat{y}$ value

Confidence interval for the mean $\hat{y}$ value, given these x's

Prediction interval for an individual $\hat{y}$ value, given these x's

Residuals in Multiple Regression

Two variable model



Nonlinear Regression Models

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

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon}$$

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

Quadratic Regression Model

Model form:

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

- where:

  β_0 = Y intercept

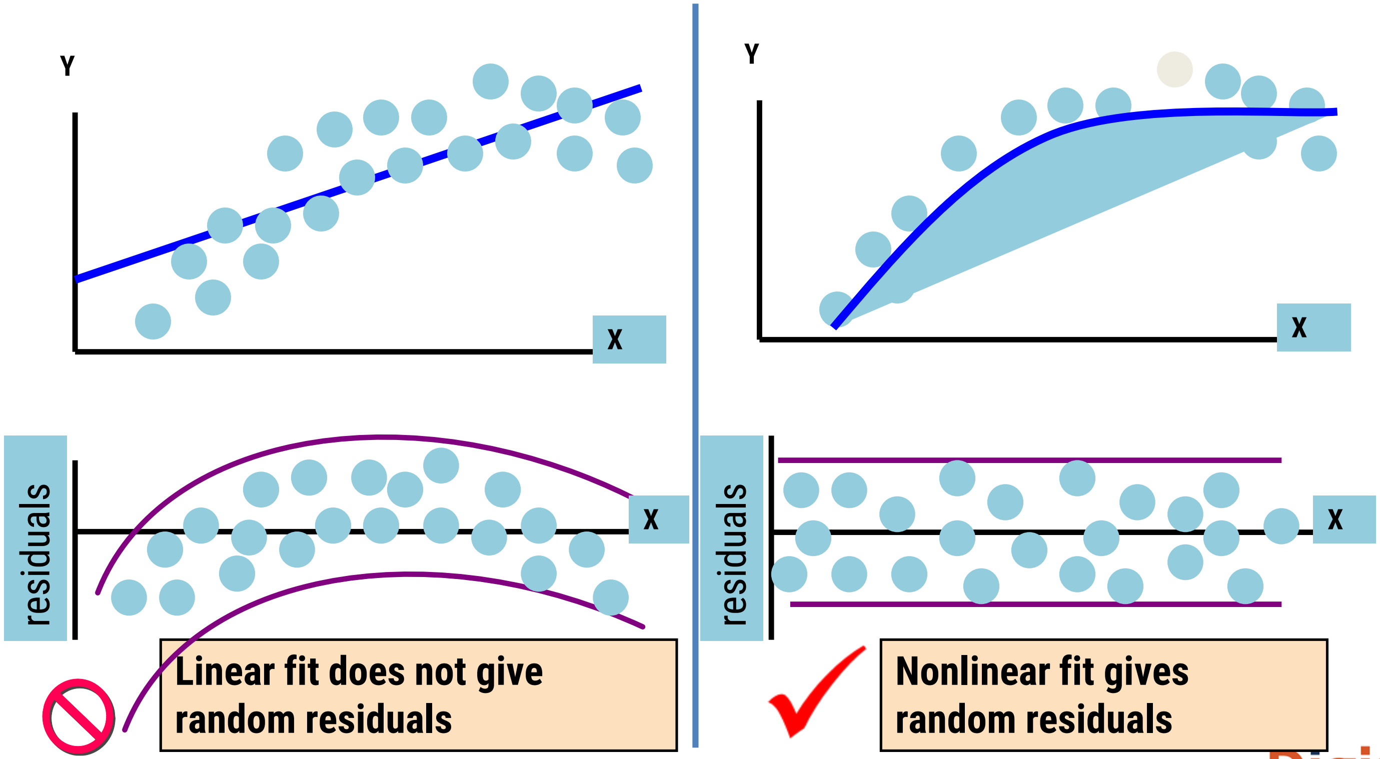
  β_1 = regression coefficient for linear effect of X on Y

  β_2 = regression coefficient for quadratic effect on Y

  ε_i = random error in Y for observation i

Digital Vidya

Linear vs. Nonlinear Fit

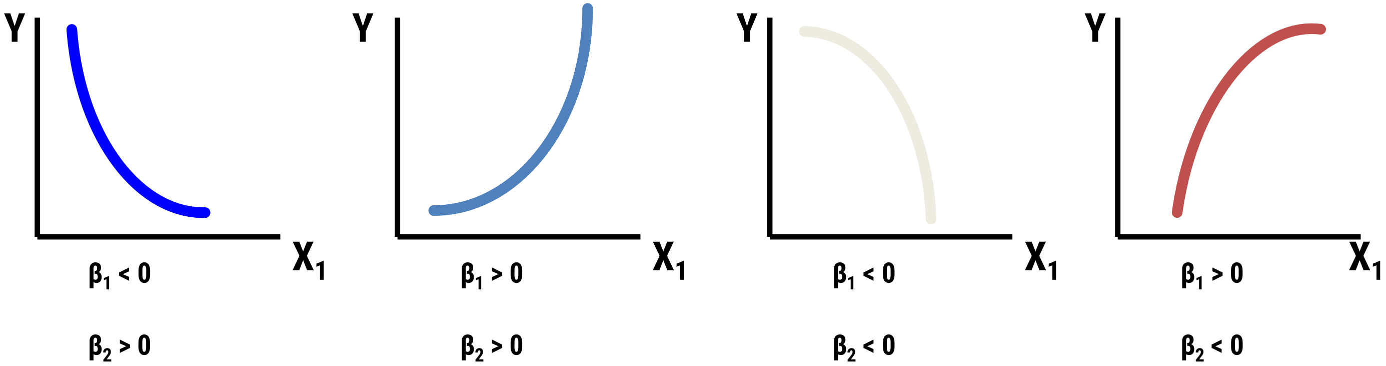


Digital Vidya

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 diagram takes on one of the following shapes:



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

Testing for Significance: Quadratic Effect

- Testing the Quadratic Effect
  - Compare the linear regression estimate

$$\hat{y} = b_0 + b_1 x_1$$

  - with quadratic regression estimate

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

  - Hypotheses

H_0: $\beta_2 = 0$ (The quadratic term does not improve the model)

H_1: $\beta_2 \neq 0$ (The quadratic term improves the model)

Testing for Significance: Quadratic Effect

(continued)

- Testing the Quadratic Effect Hypotheses
  - $H_0: \beta_2 = 0$ (The quadratic term does not improve the model)
  - $H_1: \beta_2 \neq 0$ (The quadratic term improves the model)
- The test statistic is

$$\boxed{t = \frac{b_2 - \beta_2}{s_{b_2}}}$$

$$d.f. = n - 3$$

where:

b_2 = squared term slope coefficient

β_2 = hypothesized slope (zero)

S_{b_2} = standard error of the slope

Testing for Significance: Quadratic Effect

(continued)

- Testing the Quadratic Effect

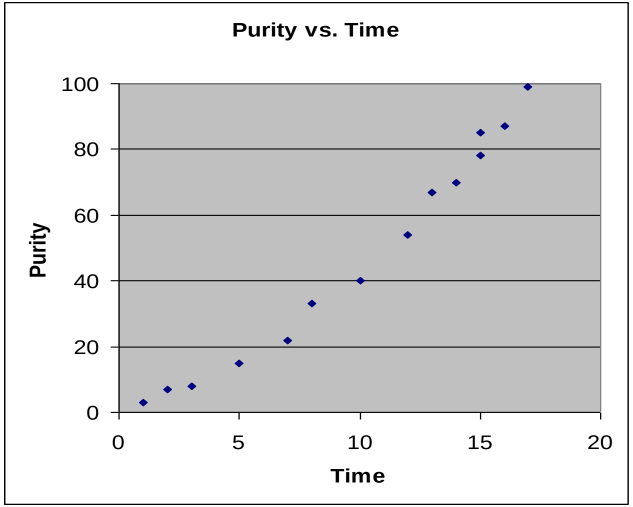
Compare R^2 from simple regression to
R^2 from the quadratic model

- If R^2 from the quadratic model is larger than R^2 from the simple model, then the quadratic model is a better model

Example: Quadratic Model

Purity	Filter Time
3	1
7	2
8	3
15	5
22	7
33	8
40	10
54	12
67	13
70	14
78	15
85	15
87	16
99	17

- Purity increases as filter time increases:



Example: Quadratic Model

(continued)

- Simple regression results:

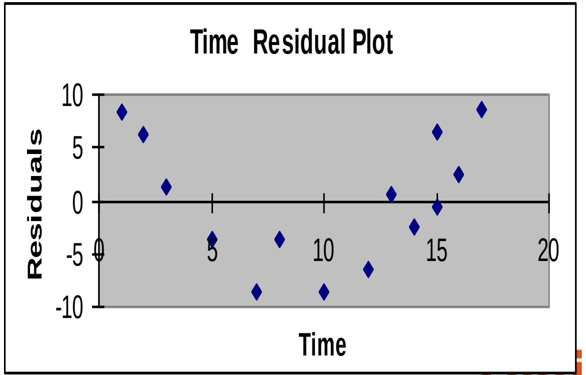
$\hat{y} = -11.283 + 5.985$ Time

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-11.28267	3.46805	-3.25332	0.00691
Time	5.98520	0.30966	**19.32819**	2.078E-10

Regression Statistics	
R Square	**0.96888**
Adjusted R Square	0.96628
Standard Error	**6.15997**

F	*Significance F*
373.57904	2.0778E-10

t statistic, F statistic, and R^2 are all high, but the residuals are not random:



Example: Quadratic Model

(continued)

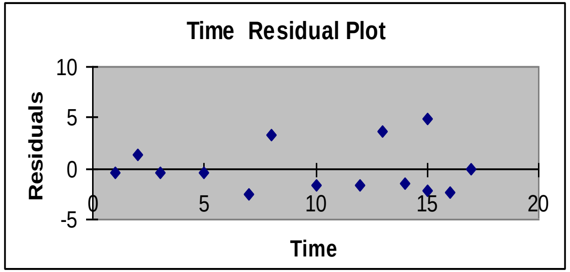
- Quadratic regression results:

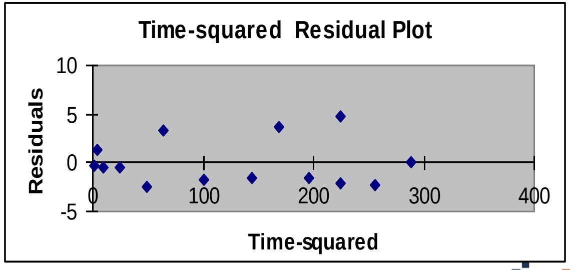
$$\hat{y} = 1.539 + 1.565\ \text{Time} + 0.245\ (\text{Time})^2$$

	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	1.53870	2.24465	0.68550	0.50722
Time	1.56496	0.60179	2.60052	0.02467
Time-squared	0.24516	0.03258	**7.52406**	1.165E-05

Regression Statistics	
R Square	0.99494
Adjusted R Square	**0.99402**
Standard Error	**2.59513**

F	***Significance F***
1080.7330	2.368E-13





The quadratic term is significant and improves the model: $\bar{R}^2$ is higher and s_e is lower, residuals are now random

Digital Vidya

The Log Transformation

The Multiplicative Model:

- Original multiplicative model

$$\boxed{Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon}$$

- Transformed multiplicative model

$$\boxed{\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \log(\varepsilon)}$$

Interpretation of coefficients

For the multiplicative model:

$$\boxed{\log Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \log \varepsilon_i}$$

- When both dependent and independent variables are logged:
  - The coefficient of the independent variable X_k can be interpreted as a 1 percent change in X_k leads to an estimated b_k percentage change in the average value of Y

  - b_k is the elasticity of Y with respect to a change in X_k

Dummy Variables

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

Dummy Variable Example

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

Let:

y = Pie Sales

x_1 = Price

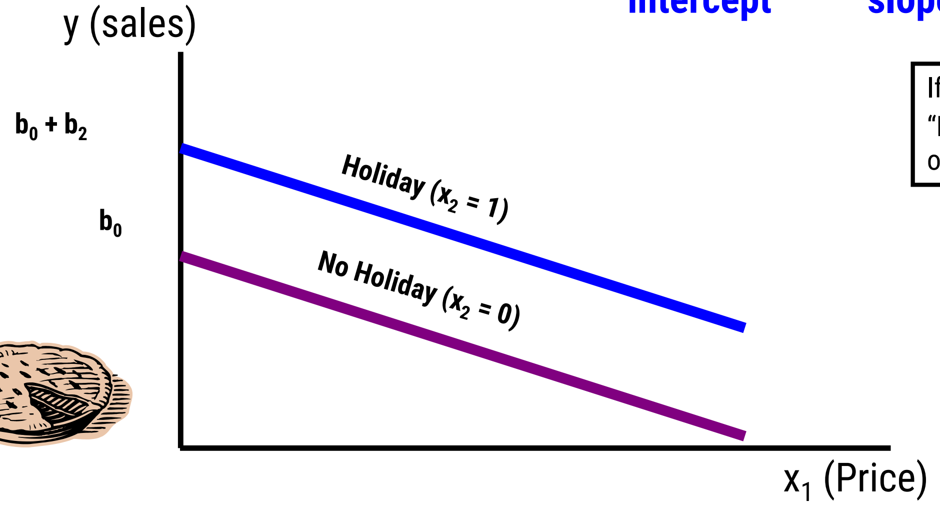
x_2 = Holiday (X_2 = 1 if a holiday occurred during the week)
(X_2 = 0 if there was no holiday that week)

Dummy Variable Example

(continued)

$\hat{y} = b_0 + b_1 x_1 + b_2(1) = (b_0 + b_2) + b_1 x_1$	**Holiday**
$\hat{y} = b_0 + b_1 x_1 + b_2(0) = b_0 + b_1 x_1$	**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 Dummy Variable Coefficient

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

Sales: number of pies sold per week
Price: pie price in \$

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

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

Digital Vidya

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

Analysis of Residuals 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

Problems and Pitfalls of Applying Least Squares Regression

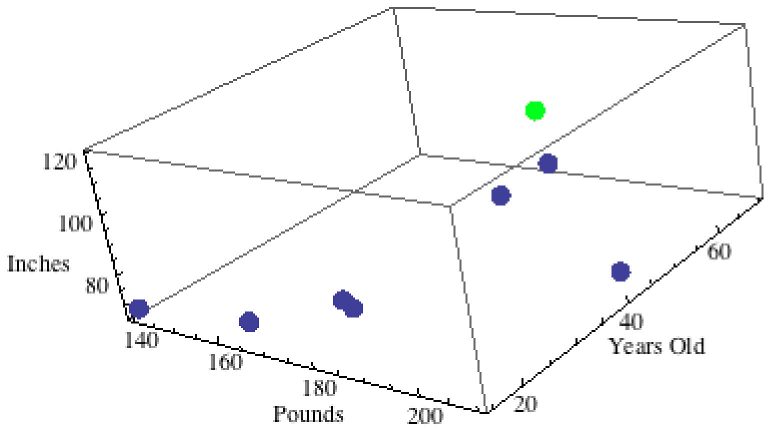
Outliers

- Least squares method is concerned with minimizing the sum of the squared error, any training point that has a dependent value that differs a lot from the rest of the data will have a disproportionately large effect on the resulting constants that are being solved for.
- Due to the squaring effect of least squares, a person in our training set whose height is mis-predicted by four inches will contribute sixteen times more error to the summed of squared errors that is being minimized than someone whose height is mis-predicted by one inch.
- That means that the more abnormal a training point's dependent value is, the more it will alter the least squares solution.
- If the outlier is sufficiently bad, the value of all the points besides the outlier will be almost completely ignored merely so that the outlier's value can be predicted accurately.

Problems and Pitfalls of Applying Least Squares Regression

Outliers

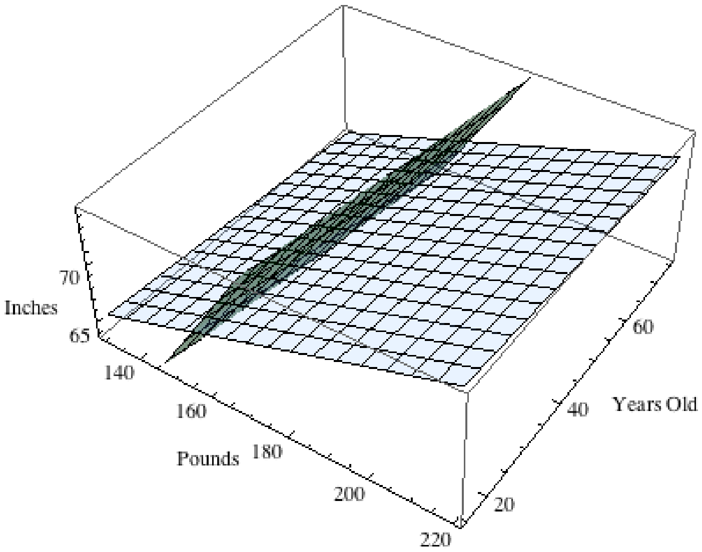
Here we see a plot of sample training data set (in purple) together with an outlier point (in green):



Problems and Pitfalls of Applying Least Squares Regression

Outliers

Below we have a plot of the old least squares solution (in blue) prior to adding the outlier point to our training set, and the new least squares solution (in green) which is attained after the outlier is added:



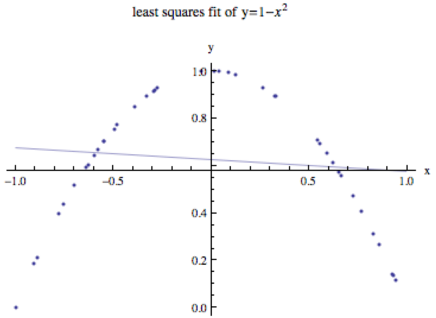
Outlier we added dramatically distorts the least squares solution and hence will lead to much less accurate predictions

Problems and Pitfalls of Applying Least Squares Regression

Non-Linearities

All linear regression methods (including, of course, least squares regression), suffer from the major drawback that in reality most systems are not linear.

Real world relationships tend to be more complicated than simple lines or planes, meaning that even with an infinite number of training points (and hence perfect information about what the optimal choice of plane is) linear methods will often fail to do a good job at making predictions



Notice that the least squares solution line does a terrible job of modelling the training points.

Digital Vidya

Problems and Pitfalls of Applying Least Squares Regression

Multi-collinearity

Multi-collinearity is a statistical phenomenon in which multiple independent variables show high correlation between each other. In other words, the variables used to predict the independent one are too inter-related.

Multi-collinearity has different causes: one of the most common is the inclusion of variables that result from mathematical operations between two or more of the other variables in the model,

e.g. net profit, which is computed by deducting total expenses from total revenues. Also, if the same kind of variable is used for the model, collinearity will always appear e.g. if you are measuring sales in both units and monetary figures the variable has the same kind.

Problems and Pitfalls of Applying Least Squares Regression

Heteroscedasticity

Heteroscedasticity refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

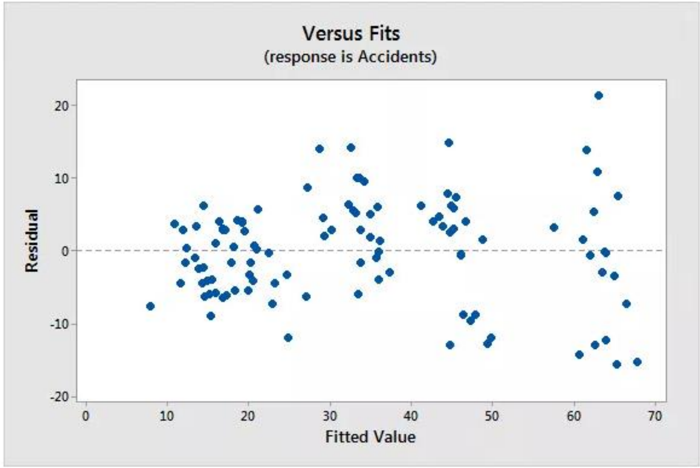
A scatterplot of these variables will often create a cone-like shape, as the scatter (or variability) of the dependent variable (DV) widens or narrows as the value of the independent variable (IV) increases. The inverse of heteroscedasticity is homoscedasticity, which indicates that a DV's variability is equal across values of an IV.

Heteroscedasticity produces a distinctive fan or cone shape in residual plots. To check for heteroscedasticity, we need to assess the residuals by fitted value plots specifically. Typically, the pattern for heteroscedasticity is that as the fitted values increases, the variance of the residuals also increases.

Problems and Pitfalls of Applying Least Squares Regression

Heteroscedasticity

You can see an example of this cone shaped pattern in the residuals by fitted value plot below. Note how the vertical range of the residuals increases as the fitted values increases.



Digital Vidya

Problems and Pitfalls of Applying Least Squares Regression

Heteroscedasticity

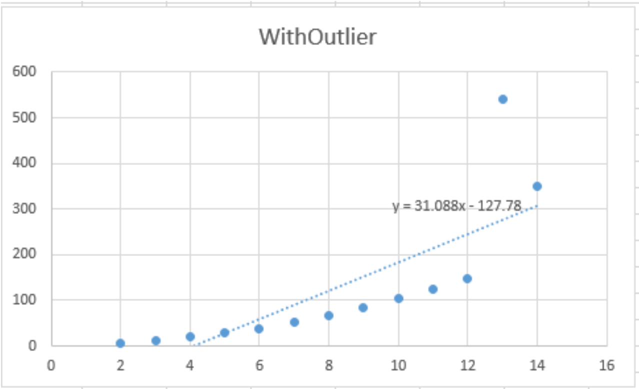
- While heteroscedasticity does not cause bias in the coefficient estimates, it does make them less precise. Lower precision increases the likelihood that the coefficient estimates are further from the correct population value.
- Heteroscedasticity tends to produce p-values that are smaller than they should be. This effect occurs because heteroscedasticity increases the variance of the coefficient estimates but the OLS procedure does not detect this increase. Consequently, OLS calculates the t-values and F-values using an underestimated amount of variance. This problem can lead you to conclude that a model term is statistically significant when it is actually not significant.

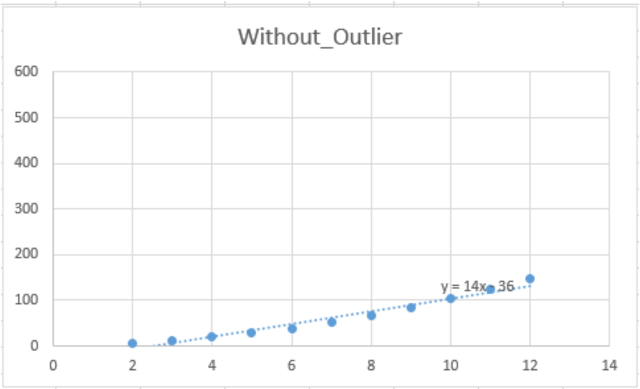
Problems and Pitfalls of Applying Least Squares Regression

Outliers

Outliers can have a dramatic impact on linear regression. It can change the model equation completely i.e. bad prediction or estimation.

Scatter plot + Linear equation with and without outlier





Digital Vidya

Problems and Pitfalls of Applying Least Squares Regression

Impact of Outliers

Outliers can drastically change the results of the data analysis and statistical modelling.

There are numerous unfavourable impacts of outliers in the data set:

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

Problems and Pitfalls of Applying Least Squares Regression

How to detect Outliers?

Most commonly used method to detect outliers is visualization. We can use various visualization methods, like Box-plot, Histogram, Scatter Plot.

Thumb rules to detect outliers:

- Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR
- Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier
- Data points, three or more standard deviation away from mean are considered outlier
- Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding
- Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. Popular indices such as Mahalanobis' distance and Cook's D are frequently used to detect outliers.

Problems and Pitfalls of Applying Least Squares Regression

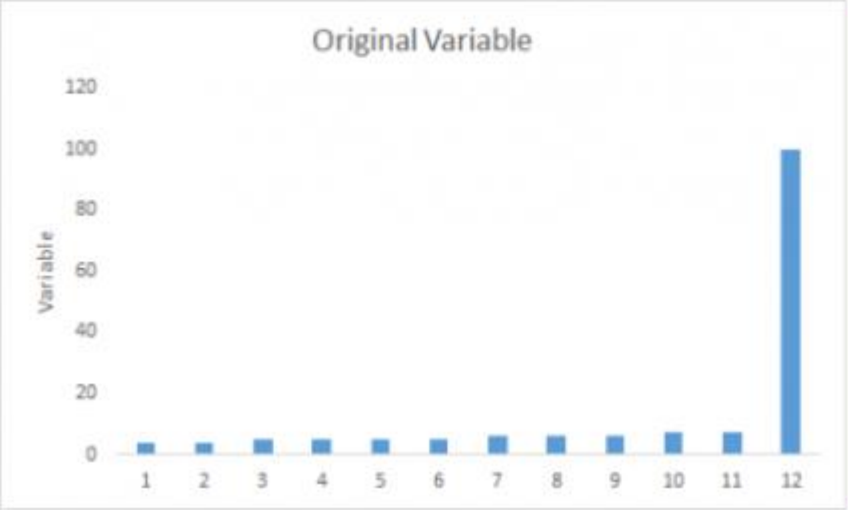
How to remove Outliers?

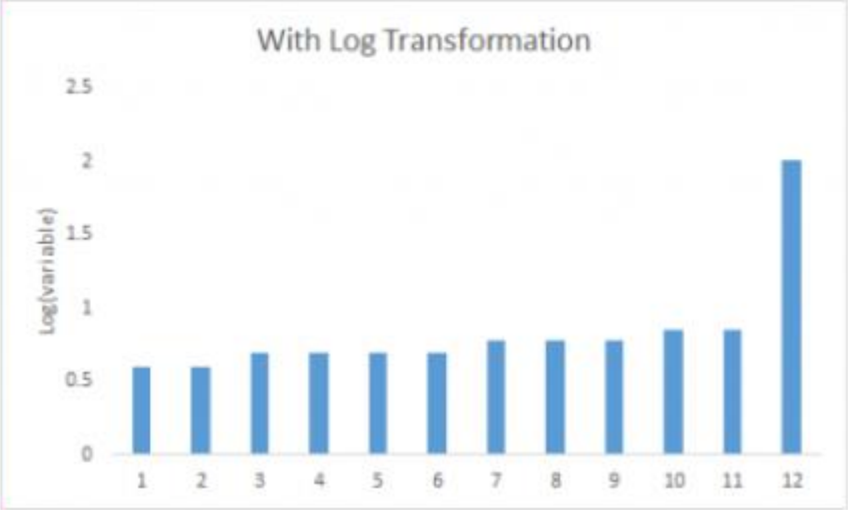
Deleting observations: We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.

Transforming and binning values: Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows to deal with outliers well due to binning of variable. We can also use the process of assigning weights to different observations.

Problems and Pitfalls of Applying Least Squares Regression

How to remove Outliers?





Variable Transformation, LOG

Problems and Pitfalls of Applying Least Squares Regression

How to remove Outliers?

Imputing:

Like imputation of missing values, we can also impute outliers. We can use mean, median, mode imputation methods. Before imputing values, we should analyse if it is natural outlier or artificial. If it is artificial, we can go with imputing values. We can also use statistical model to predict values of outlier observation and after that we can impute it with predicted values.

Thank You

Digital Vidya