Statistics Foundation

Simple Linear Regression

Digital Vidya

Correlation Analysis

- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation

Correlation Analysis

- The population correlation coefficient is denoted ρ (the Greek letter rho)
- The sample correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Digital Vidya

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

Linear Regression Model

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X
- Linear regression population equation model

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

Simple Linear Regression Model

The population regression model:

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

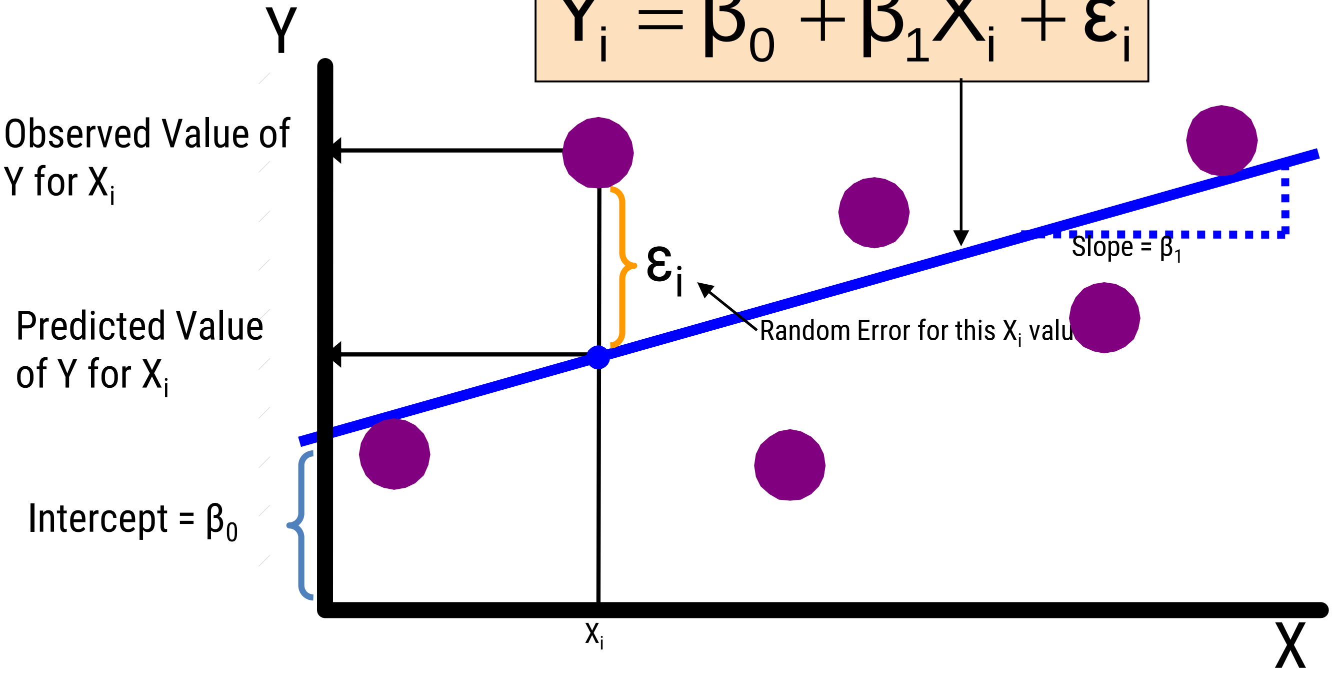
Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple Linear Regression Model

(continued)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of x for observation i

$$\hat{y}_i = b_0 + b_1 x_i$$

The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

Digital Vidya

Least Squares Estimators

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between y and $\hat{y}$:

$$\boxed{\begin{aligned} \min \ SSE &= \min \sum e_i^2 \\ &= \min \sum (y_i - \hat{y}_i)^2 \\ &= \min \sum [y_i - (b_0 + b_1 x_i)]^2 \end{aligned}}$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

Least Squares Estimators

(continued)

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_Y}{s_X}$$

- And the constant or y-intercept is

$$b_0 = \bar{y} - b_1\bar{x}$$

- The regression line always goes through the mean $\bar{x}, \bar{y}$

Finding the Least Squares Equation

- The coefficients b_0 and b_1 , and other regression results in this module, will be found using a computer
  - Hand calculations are tedious
  - Statistical routines are built into Excel
  - Other statistical analysis software can be used

Linear Regression Model Assumptions

- The true relationship form is linear (Y is a linear function of X, plus random error)
- The error terms, ε_i are independent of the x values
- The error terms are random variables with mean 0 and constant variance, σ^2

(the constant variance property is called homoscedasticity)

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

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

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

Digital Vidya

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero (if $x = 0$ is in the range of observed x values)

- b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Digital Vidya

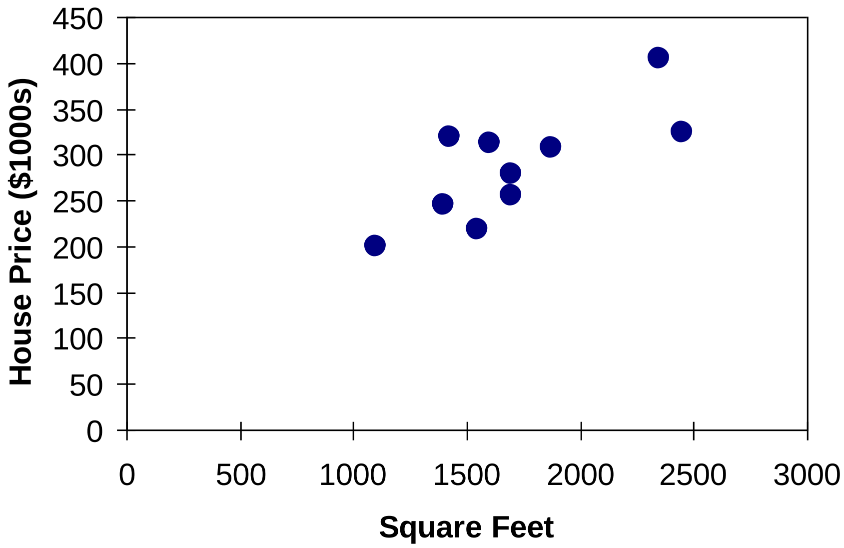
Sample Data for House Price Model

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Digital Vidya

Graphical Presentation

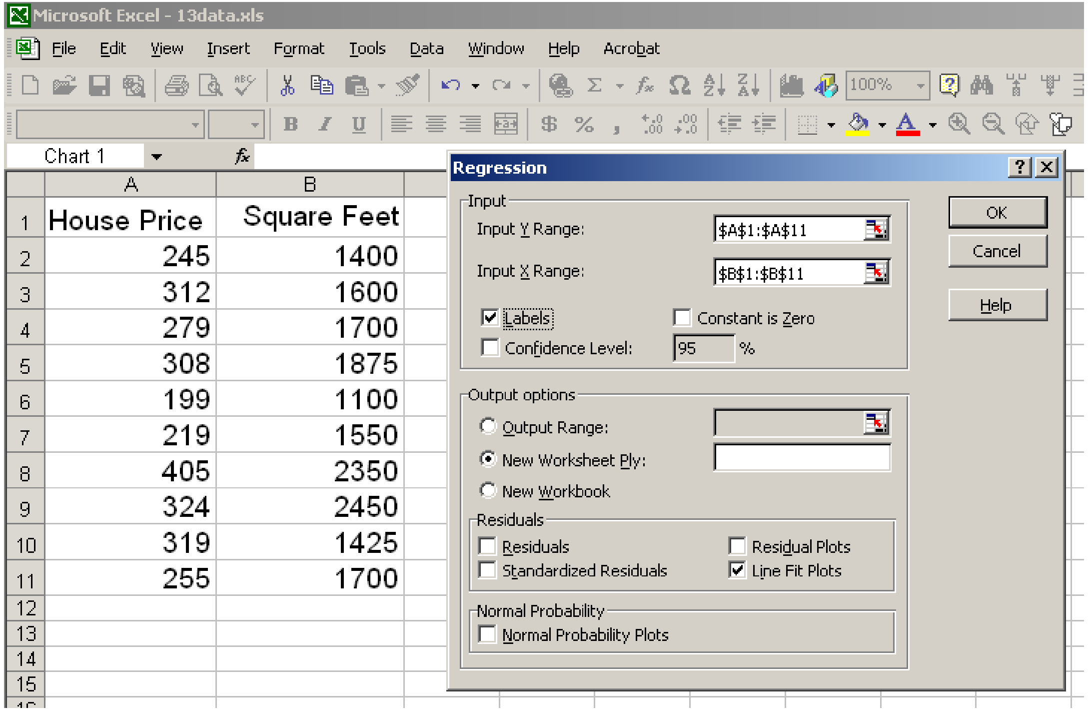
- House price model: scatter plot



Digital Vidya

Regression Using Excel

- Tools / Data Analysis / Regression



Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

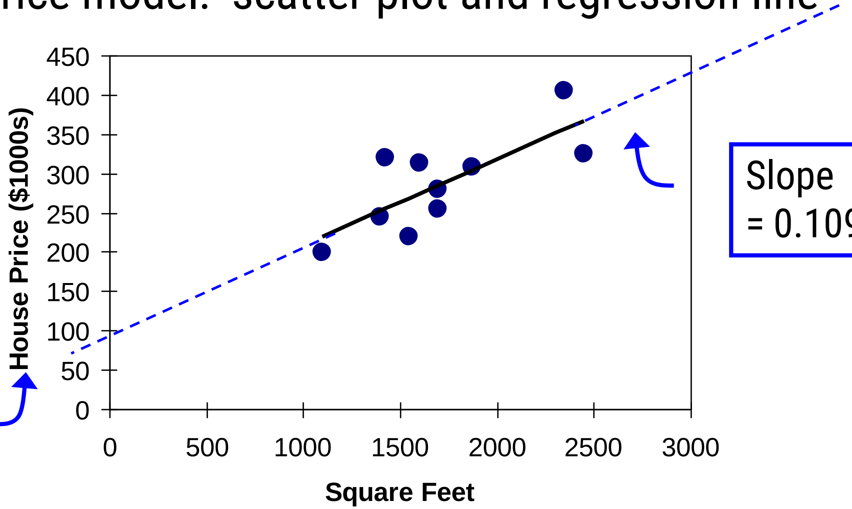
$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Graphical Presentation

- House price model: scatter plot and regression line



Intercept
= 98.248

Slope
= 0.10977

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
  - Here, no houses had 0 square feet, so $\boxed{b_0 = 98.24833}$ just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Measures of Variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (y_i - \bar{y})^2$	$SSR = \sum (\hat{y}_i - \bar{y})^2$	$SSE = \sum (y_i - \hat{y}_i)^2$

where:

$\bar{y}$ = Average value of the dependent variable

y_i = Observed values of the dependent variable

$\hat{y}_i$ = Predicted value of y for the given x_i value

Digital Vidya

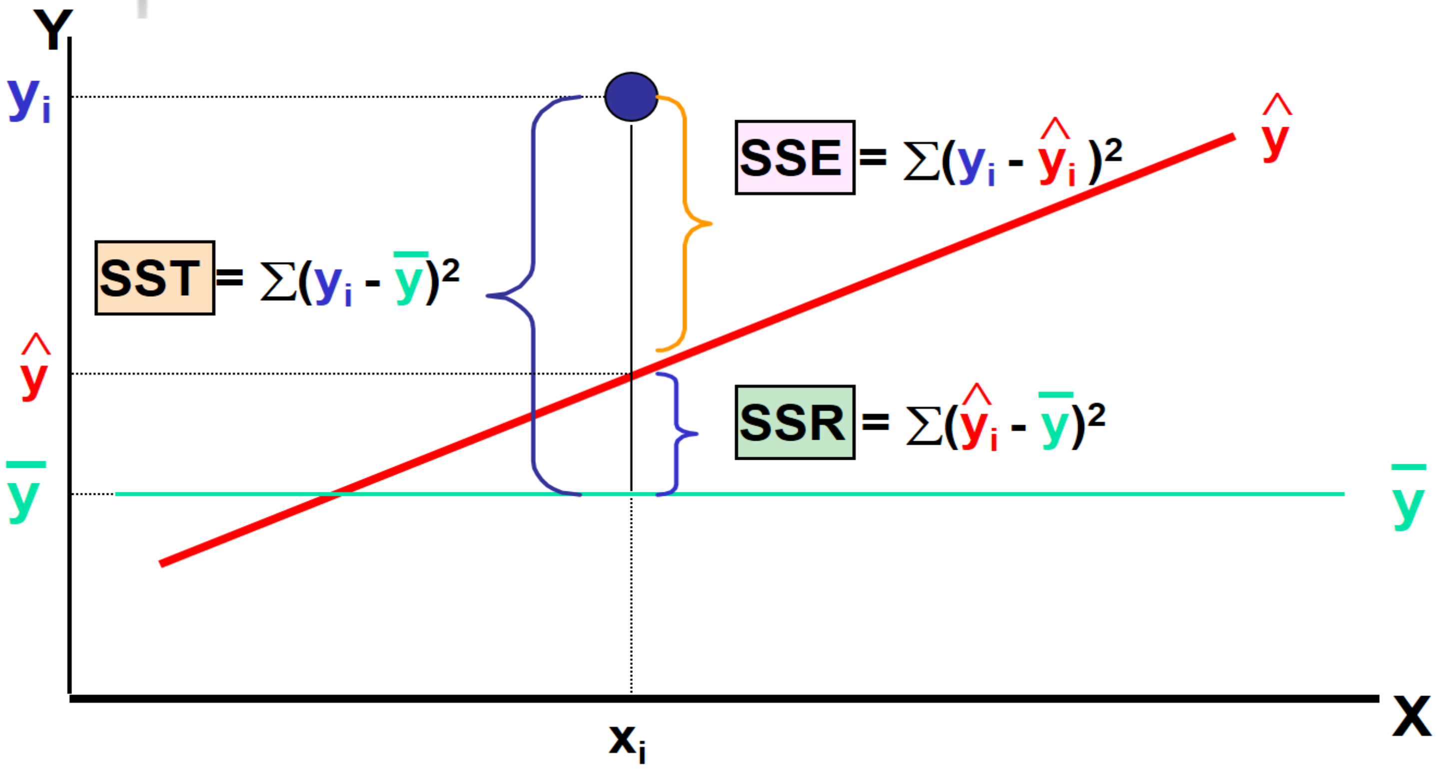
Measures of Variation

(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean, $\bar{y}$
- SSR = regression sum of squares
  - Explained variation attributable to the linear relationship between x and y
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between x and y

Measures of Variation

(continued)



Coefficient of Determination, R^2

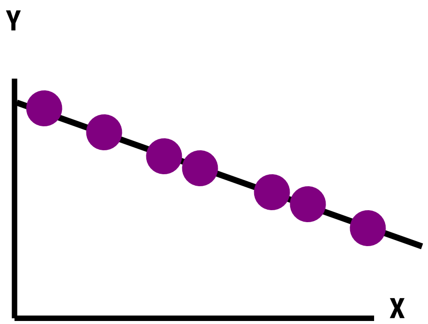
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

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

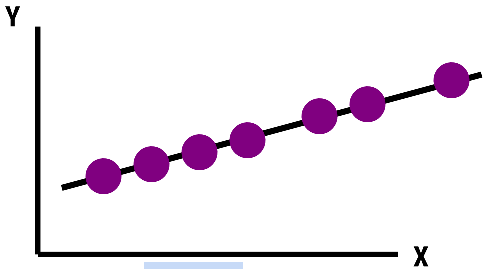
note: $0 \le R^2 \le 1$

Digital Vidya

Examples of Approximate r^2 Values



$r^2 = 1$



$r^2 = 1$

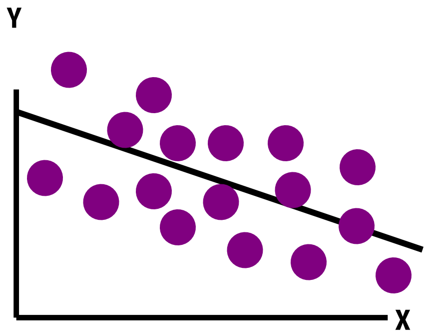
$r^2 = 1$

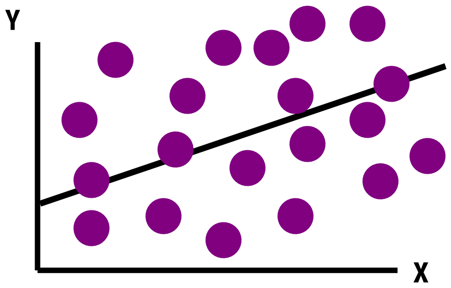
Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Digital Vidya

Examples of Approximate r2 Values





$0 < r^2 < 1$

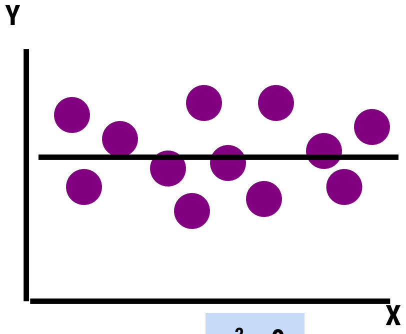
Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

Digital Vidya

Examples of Approximate r^2 Values

$r^2 = 0$



$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Digital Vidya

Correlation and R_2

- The coefficient of determination, R^2, for a simple regression is equal to the simple correlation squared

$$\boxed{R^2 = r_{xy}^2}$$

Digital Vidya

Estimation of Model Error Variance

- An estimator for the variance of the population model error is

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

- Division by n – 2 instead of n – 1 is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$$s_e = \sqrt{s_e^2}$$ is called the standard error of the estimate

Digital Vidya

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$s_e = 41.33032$$

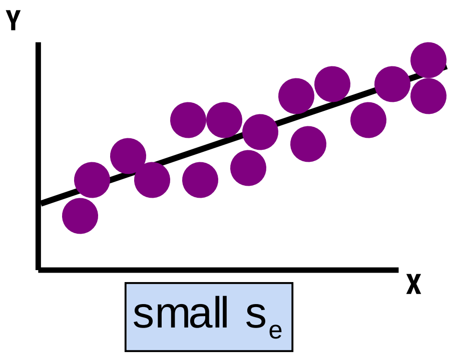
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

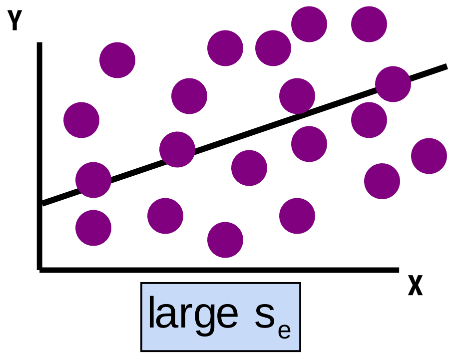
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Digital Vidya

Comparing Standard Errors

s_e is a measure of the variation of observed y values from the regression line





The magnitude of s_e should always be judged relative to the size of the y values in the sample data

i.e., s_e = \$41.33K is moderately small relative to house prices in the \$200 - \$300K range

Digital Vidya

Inferences About the Regression Model

- The variance of the regression slope coefficient (b_1) is estimated by

$$\boxed{s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2}}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_e = \sqrt{\dfrac{SSE}{n-2}}$ = Standard error of the estimate

Digital Vidya

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$s_{b_1} = 0.03297$$

ANOVA

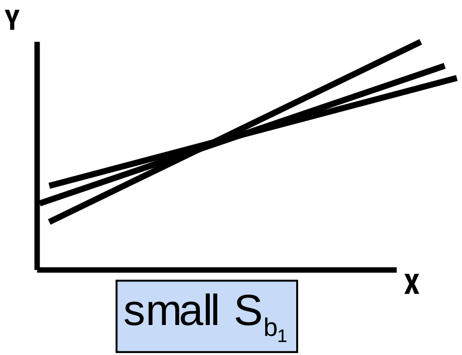
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

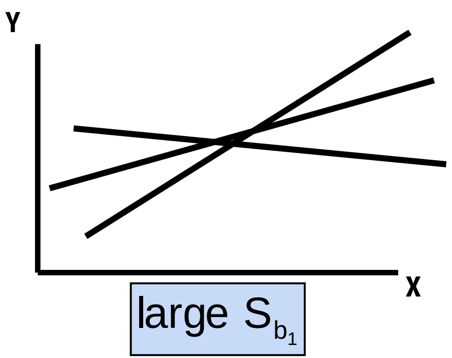
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Digital Vidya

Comparing Standard Errors of the Slope

S_{b_1} is a measure of the variation in the slope of regression lines from different possible samples





Digital Vidya

Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses

H_0:	$\beta_1 = 0$	(no linear relationship)
H_1:	$\beta_1 \neq 0$	(linear relationship does exist)

- Test statistic

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

s_{b1} = standard error of the slope

Digital Vidya

Inference about the Slope: t Test

(continued)

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house affect its sales price?

Digital Vidya

Inferences about the Slope: t Test Example

H_0: $\beta_1 = 0$
H_1: $\beta_1 <> 0$

From Excel output:

b_1 s_{b_1}

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Digital Vidya

Inferences about the Slope: t Test Example

(continued)

Test Statistic: **t = 3.329**

H_0: $\beta_1 = 0$

H_1: $\beta_1 <> 0$

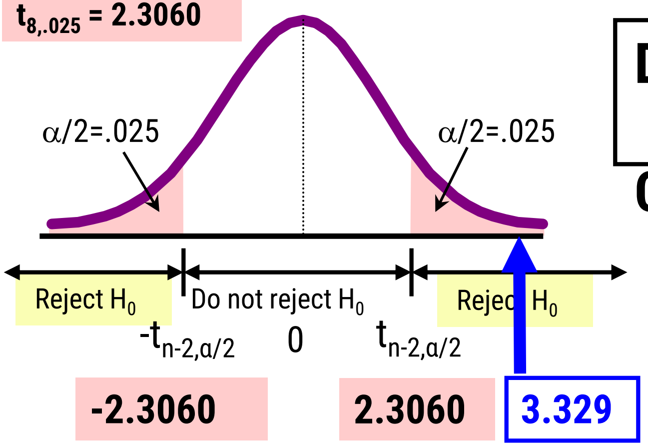
From Excel output:

b_1 S_{b_1} t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

d.f. = 10-2 = 8

$t_{8,.025}$ = 2.3060



Decision:
Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Inferences about the Slope: t Test Example

(continued)

P-value = **0.01039**

H_0: $\beta_1 = 0$

H_1: $\beta_1 <> 0$

From Excel output:

P-value

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

This is a two-tail test, so the p-value is

$P(t > 3.329)+P(t < -3.329) = 0.01039$

(for 8 d.f.)

Decision: P-value < α so Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Digital Vidya

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 - t_{n-2,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2} s_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Confidence Interval Estimate for the Slope

(continued)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Digital Vidya

F-Test for Significance

- F Test statistic: $$F = \frac{MSR}{MSE}$$

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

where F follows an F distribution with k numerator and (n – k - 1) denominator degrees of freedom

(k = the number of independent variables in the regression model)

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

P-value for the F-Test

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Digital Vidya

F-Test for Significance

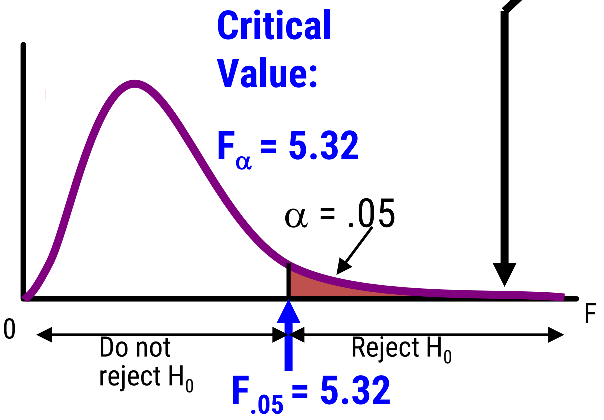
(continued)

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$\alpha = .05$

$\underbrace{df_1 = 1 \qquad df_2 = 8}$



Test Statistic:

$$F = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Prediction

- The regression equation can be used to predict a value for y, given a particular x

- For a specified value, x_{n+1} , the predicted value is

$$\boxed{\hat{y}_{n+1} = b_0 + b_1 x_{n+1}}$$

Predictions Using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

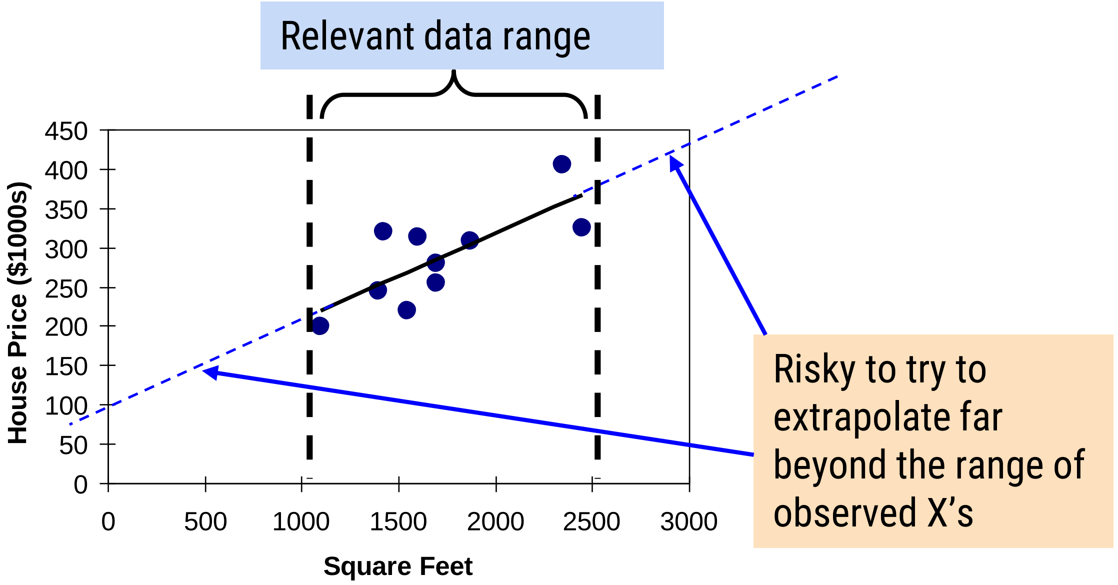
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Relevant Data Range

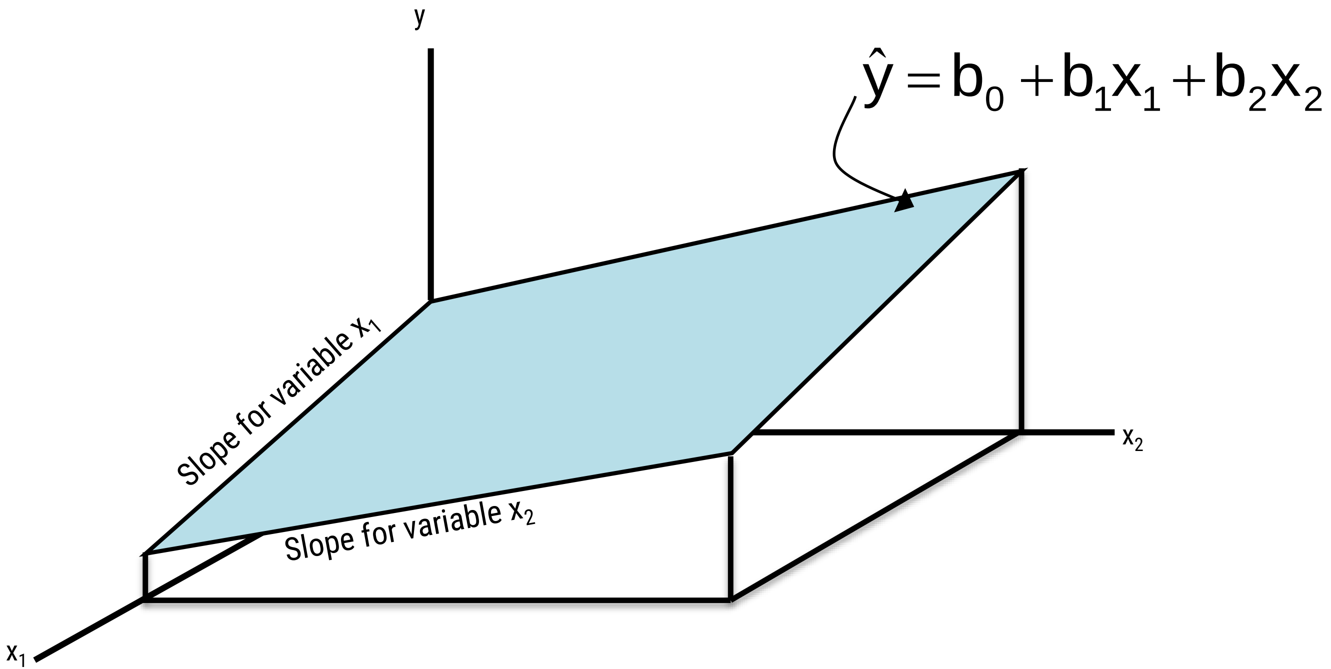
- When using a regression model for prediction, only predict within the relevant range of data



Multiple Regression Equation

(continued)

Two variable model



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.

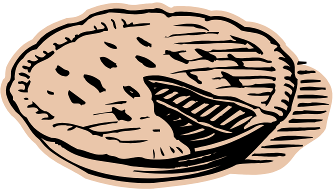
$$\boxed{E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \qquad \text{for } (i = 1, \ldots, n)}$$

(The constant variance property is called homoscedasticity)

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)



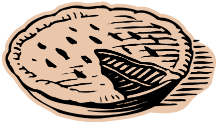
- Data is collected for 15 weeks

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:

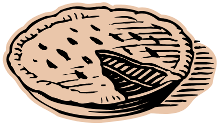
$$\widehat{\text{Sales}} = b_0 + b_1 (\text{Price}) + b_2 (\text{Advertising})$$



Digital Vidya

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

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

- The unbiased estimate of the variance of the errors is

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

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

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

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

(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

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

- 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

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)

Digital Vidya

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:

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

where, $(df = n - k - 1)$

Digital Vidya

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

Digital Vidya

Example: Evaluating Individual Regression Coefficients

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

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

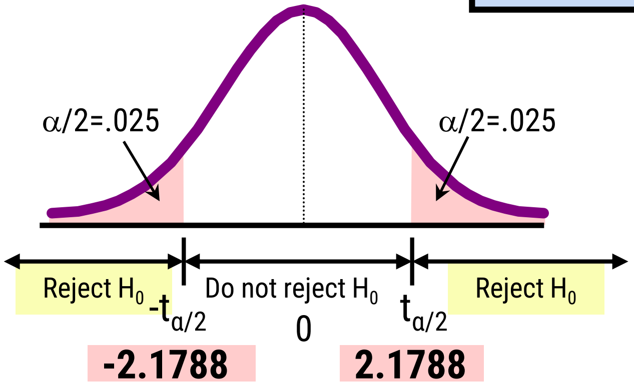
α = .05

$t_{12, .025}$ = 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: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)

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

Digital Vidya

F-Test for Overall Significance

- Test statistic:

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

where F has k (numerator) and

(n – K – 1) (denominator)

degrees of freedom

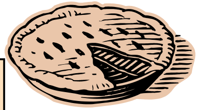
- The decision rule is

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

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)

$H_0: \beta_1 = \beta_2 = 0$

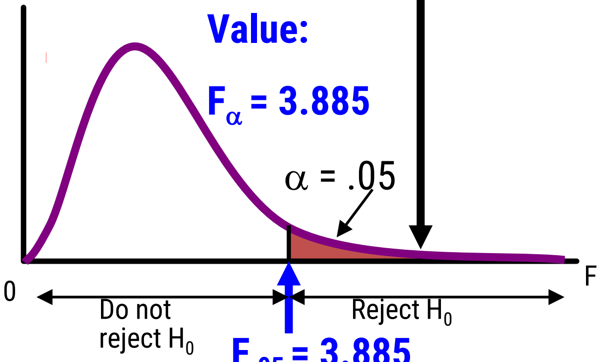
$H_1: \beta_1$ and β_2 not both zero

$\alpha = .05$

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

Critical Value:

$F_\alpha = 3.885$



$F_{.05} = 3.885$

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

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

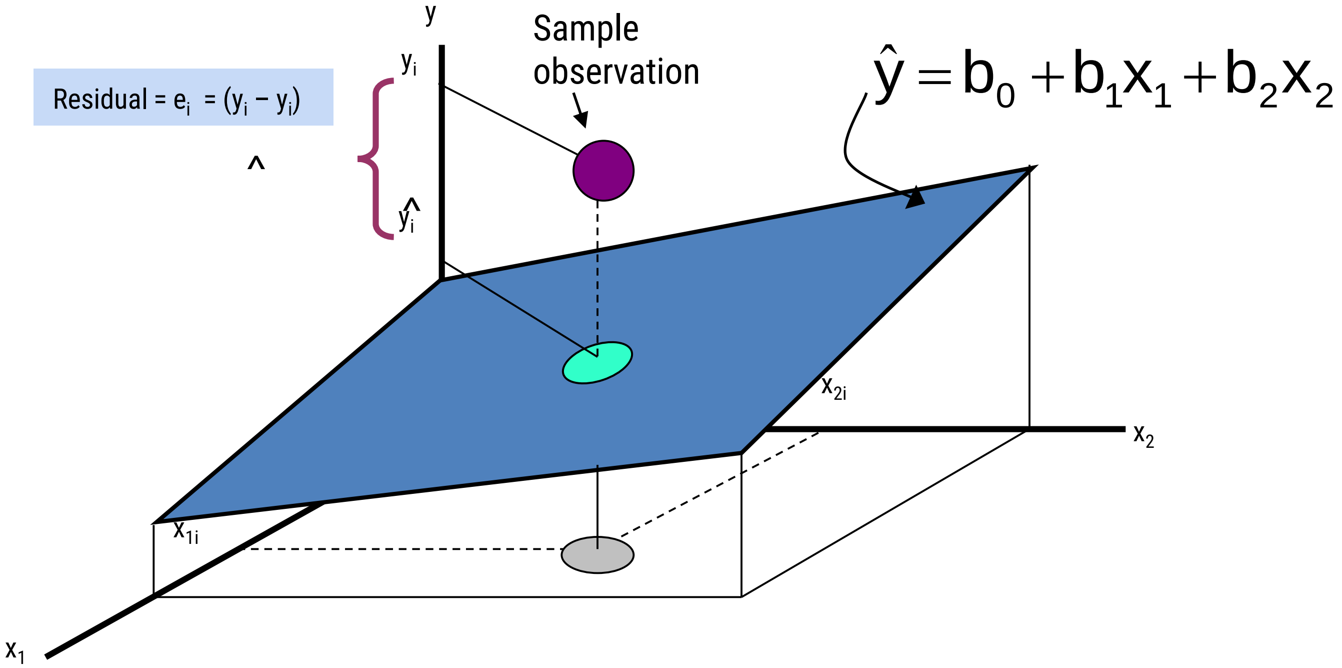
Note that Advertising is in \$100's, so \$350 means that $X_2 = 3.5$

Digital Vidya

Residuals in Multiple Regression

Two variable model

Residual = $e_i = (y_i - \hat{y}_i)$



Digital Vidya

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:

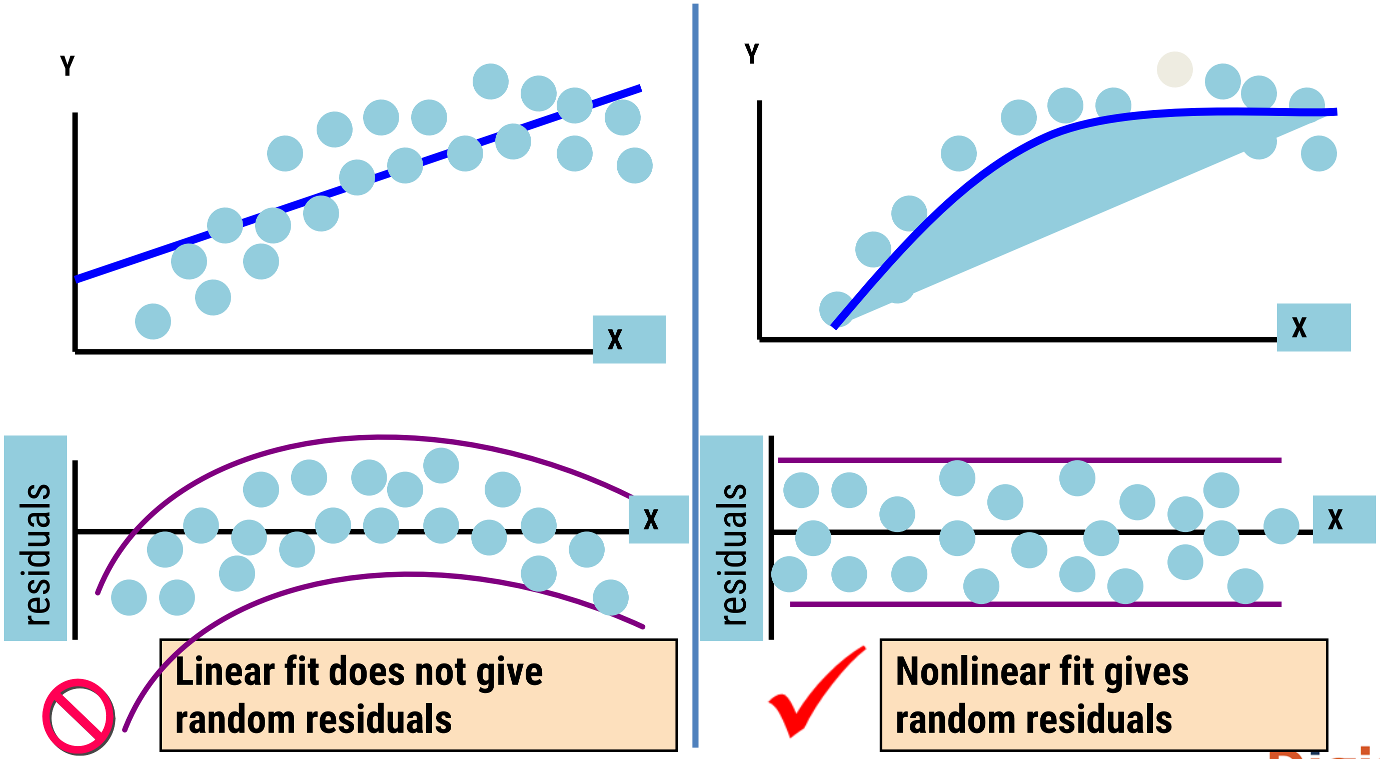
  β_0 = Y intercept

  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

Linear vs. Nonlinear Fit



Digital Vidya

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

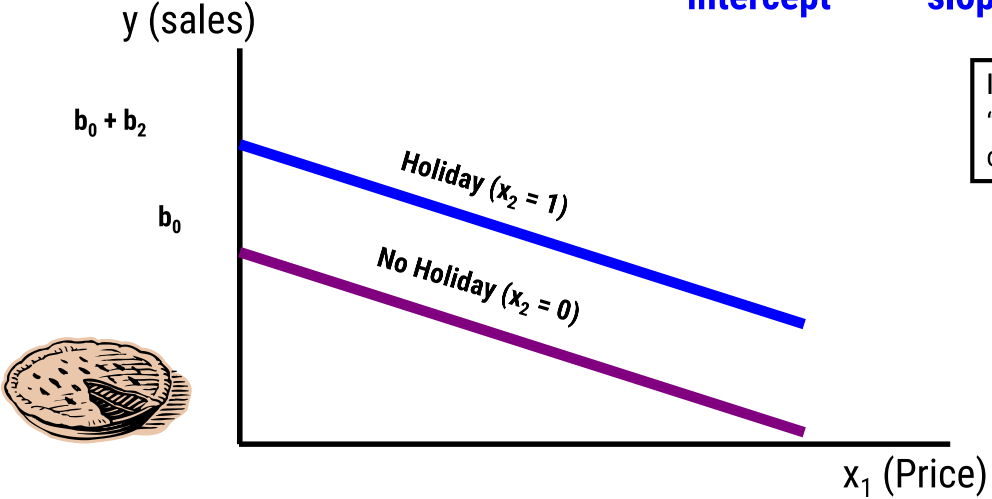
x_2 = Holiday (X_2 = 1 if a holiday occurred during the week)
(X_2 = 0 if there was no holiday that week)

Dummy Variable Example

(continued)

$\hat{y} = b_0 + b_1 x_1 + b_2(1) = (b_0 + b_2) + b_1 x_1$	**Holiday**
$\hat{y} = b_0 + b_1 x_1 + b_2(0) = \quad b_0 \quad + b_1 x_1$	**No Holiday**

Different intercept **Same slope**



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

Digital Vidya

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

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

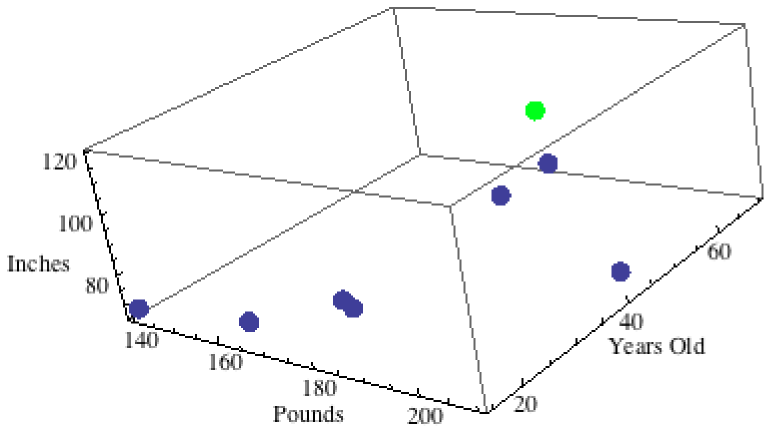
<u>Outliers</u>

- 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

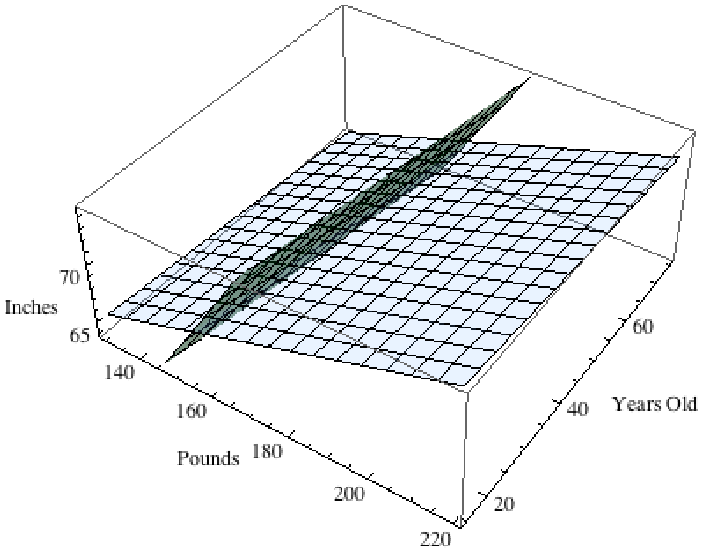
Here we see a plot of sample training data set (in purple) together with an outlier point (in green):



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:



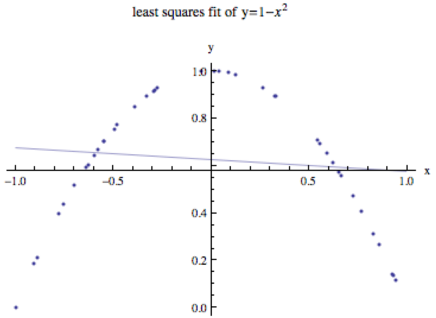
Outlier we added dramatically distorts the least squares solution and hence will lead to much less accurate predictions

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.

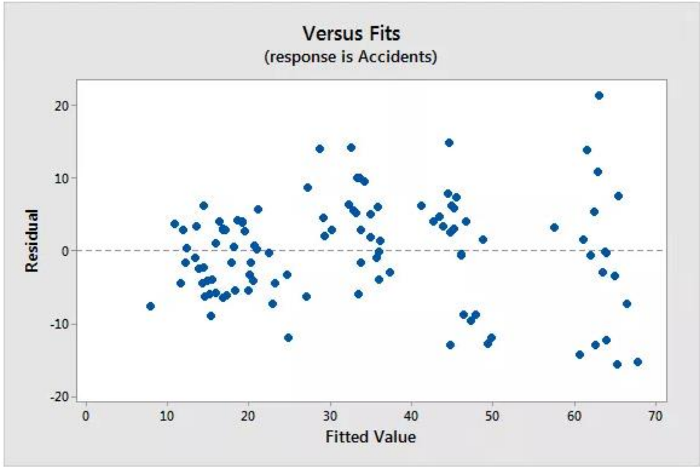
A scatterplot of these variables will often create a cone-like shape, as the scatter (or variability) of the dependent variable (DV) widens or narrows as the value of the independent variable (IV) increases. The inverse of heteroscedasticity is homoscedasticity, which indicates that a DV's variability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

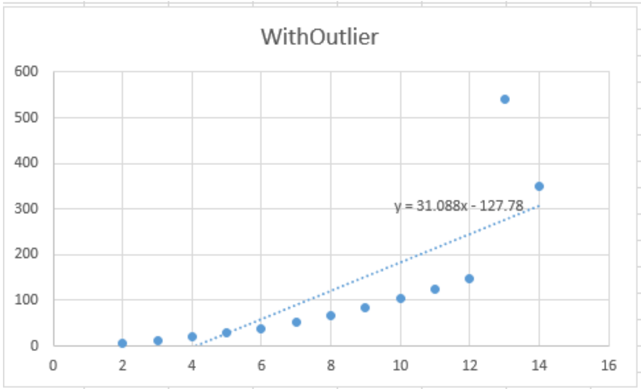
- While heteroscedasticity does not cause bias in the coefficient estimates, it does make them less precise. Lower precision increases the likelihood that the coefficient estimates are further from the correct population value.
- Heteroscedasticity tends to produce p-values that are smaller than they 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.

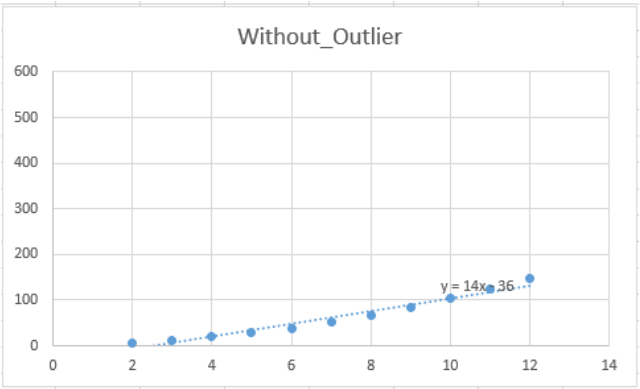
Problems and Pitfalls of Applying Least Squares Regression

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

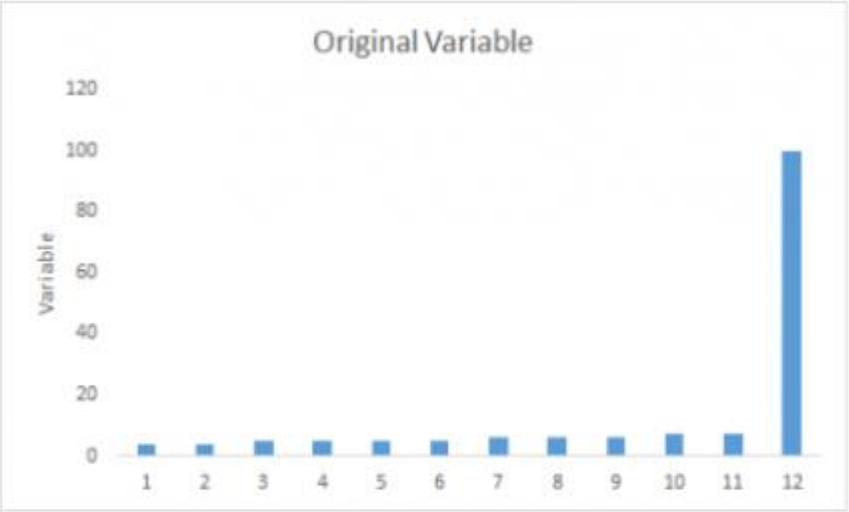
How to remove Outliers?

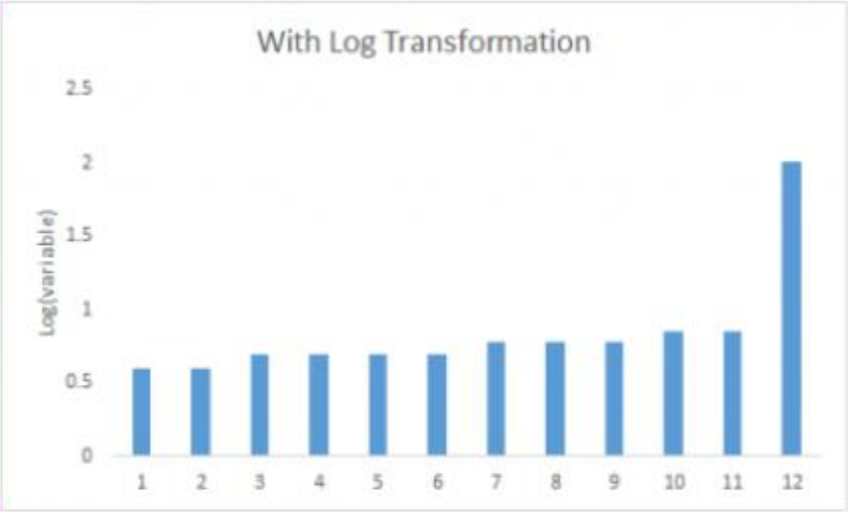
Deleting observations: We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.

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

<u>How to remove Outliers?</u>





<u>Variable Transformation, LOG</u>

Problems and Pitfalls of Applying Least Squares Regression

<u>How to remove Outliers?</u>

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.

Derivation

Direct regression method

This method is also known as the **ordinary least squares estimation**. Assuming that a set of n paired observations on (x_i, y_i), $i = 1, 2, ..., n$ are available which satisfy the linear regression model $y = \beta_0 + \beta_1 X + \varepsilon$.

So we can write the model for each observation as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $(i = 1, 2, ..., n)$.

The direct regression approach minimizes the sum of squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to β_0 and β_1.

The partial derivatives of $S(\beta_0, \beta_1)$ with respect to β_0 is

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

and the partial derivative of $S(\beta_0, \beta_1)$ with respect to β_1 is

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i .$$

Derivation

The solutions of β_0 and β_1 are obtained by setting

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

The solutions of these two equations are called the **direct regression estimators**, or usually called as the **ordinary least squares (OLS)** estimators of β_0 and β_1.

This gives the ordinary least squares estimates b_0 of β_0 and b_1 of β_1 as

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s_{xy}}{s_{xx}}$$

where

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \; \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Thank You