Chapter 11

Regression Analysis:

Regression analysis describes the relationship between two quantitative variables in a specific setting. Of the two variables, the 'variable of interest' in a study is known as the "**dependent**" variable and the other variable is called the "**independent** "variable that explains changes in the dependent variable.
The **regression line** predicts the value for the response variable y as a straight line function of the value of the explanatory variable x. This line describes how a response variable *y* changes as an explanatory variable *x* changes.

Let $\hat{y}$ (y hat) denote the predicted value of y. The equation for the simple linear regression line is given by,

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

The constant term a denotes the **y-intercept** of the equation. The value of b_o is the height of the line above the value of $x = 0$.

The constant term b_1 denotes the ***slope*** of the regression equation. The value of b_1 is the amount by which y increases when x increases by ***one*** unit.

The prediction error or ***residual*** for an observation is the difference between the actual value and the predicted value of the response variable y.

Residual = $y - \hat{y}$ = *Observed – predicted.*

The formula for computing ***slope*** is, $b_1 = r\left(\frac{s_y}{s_x}\right)$

s_x is the standard deviation of the explanatory variable x,

s_y is the standard deviation of the response variable y.

The formula for computing **y-intercept** is, $b_0 = \bar{y} - b\bar{x}$

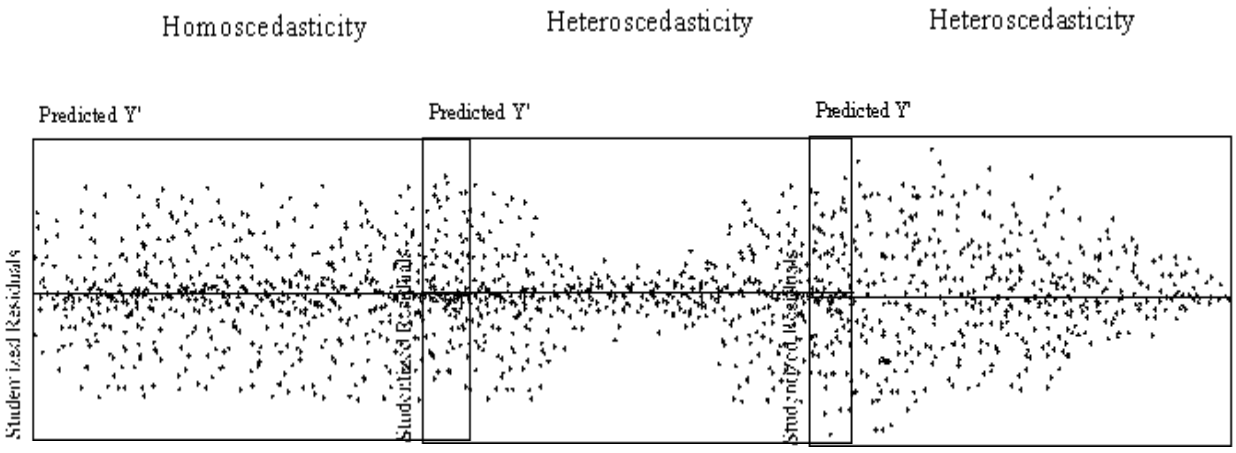
In the above equation, $\bar{y}$ is mean of y and $\bar{x}$ is the mean of x.

Proportional reduction in error (r^2) is the proportion of accuracy (in terms of explaining the variability) when regression equation is used to predict y values for each specific x values.

r-squared = r^2 = (correlation)^2.

The r^2 is also known as the **coefficient of determination**. An interpretation of r^2 =0.80 is: approximately 80% of the variability in the *response variable* can be explained by this linear regression equation.

Examples of random and non-random residual plots



Regression Analysis Example:

Example: A random sample of 9 pair of data points representing home size and price are given below. Investigate the correlation between home size and price.

Home size (in hundreds of square feet): 26, 27, 33, 29, 29, 34, 30, 40, 22
Home price (in thousands of dollar): 259, 274, 294, 296, 325, 380, 457, 523, 215.

Descriptive Statistics

	Mean	Std. Deviation	N
Home Size (in hundreds of sqft)	30.00	5.196	9
Home Price(in thousands of dollars)	335.89	99.652	9

Correlations

		Home Size (in hundreds of sqft)	Home Price(in thousands of dollars)
Home Size (in hundreds of sqft)	Pearson Correlation	1	.829**
	Sig. (2-tailed)		.006
	N	9	9
Home Price(in thousands of dollars)	Pearson Correlation	.829**	1
	Sig. (2-tailed)	.006	
	N	9	9

**. Correlation is significant at the 0.01 level (2-tailed).

1. Comment on the correlation between "House Size" and "House Price". Discuss the scatter plot.

   The correlation between "House Size" and "House Price" is 0.829, which is strong and positive. The scatter plot shows a positive trend but strength of the correlation is not obvious, because of only 9 data points.

Model	Variables Entered	Variables Removed	Method
1	Home Size (in hundreds of sq.ft.)[b]	.	Enter

a. Dependent Variable: Home Price (in thousands of dollars)

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.829[a]	.687	.642	59.621

a. Predictors: (Constant), Home Size (in hundreds of sq.ft.)

b. Dependent Variable: Home Price (in thousands of dollars)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-140.917	123.312		-1.143	.291
	Home Size (in hundreds of sq.ft.)	15.894	4.057	.829	3.918	.006

a. Dependent Variable: Home Price (in thousands of dollars)

Answer the following questions based on the SPSS regression analysis output from the previous page.

1. What are the intercept and slope of the model?
   Intercept: 160.194. Slope: 0.099.

2. Write down the regression model.
   Y-hat = 160.194 + 0.099*x*

3. What is the value of R^2? Give interpretation in the context of the problem.
   R^2 value is 0.687. Approximately 69% of the variation in "House Price" can be explained by this regression model based on "House Size".

4. Predict the price of a house (in thousands of dollars) if the house size is 2700 sq. ft.
   Y-hat = -140.194 + 15.89*x* = -140.194 + 15.89*(27)* = 288.38.
   In other words, the predicted house price will be $288,380.00.

5. If the true price for the house of 2700 sq. ft. in question 4 is $274,000 , what is the model residual or prediction error?
   The model residual is: *Y* – *Y-hat* = observed – predicted =274 – 288.38 = – 14.38.
   In this case the model is over estimating the house price by $14,380.

6. Predict the price of a house (in thousands of dollars) if the house size is 800 sq. ft.

   Y-hat = -140.194 + 15.89*x* = -140.194 + 15.89*(8)* = -13.72.

The predicted house price is - $13,720 when the size is 800 sqft.
This is an improbable result. The reason for getting this result is that the value 800 is outside the range of the x values (independent variable), the slope and intercept are not valid for size of 800.
This type prediction problem is called extrapolation.

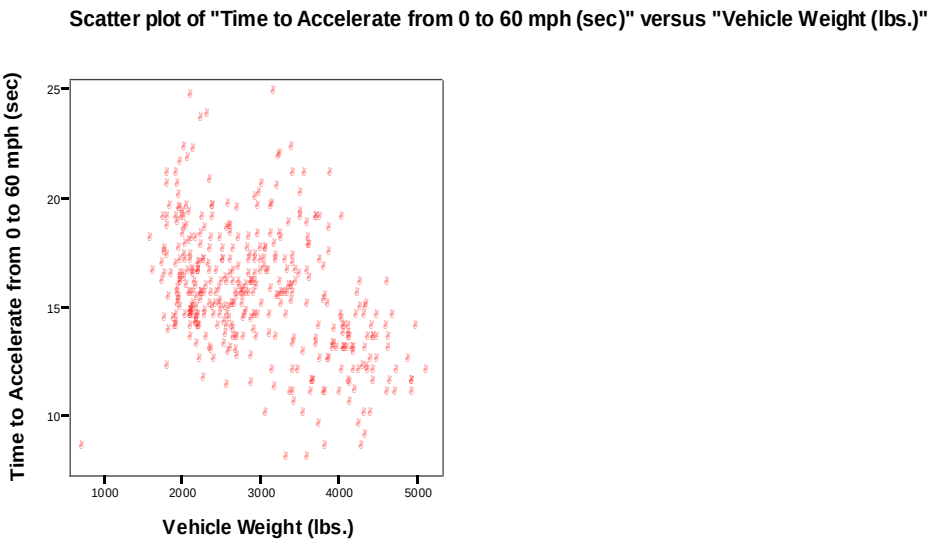
1. Describe the correlation between the two variables "*Time to Accelerate from 0 to 60 mph (sec)*" and "*Vehicle Weight (lbs.)*" from the table below.

Correlations

		Time to Accelerate from 0 to 60 mph (sec)	Vehicle Weight (lbs.)
Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	1	-.415(**)
	Sig. (2-tailed)		.000
	N	406	406
Vehicle Weight (lbs.)	Pearson Correlation	-.415(**)	1
	Sig. (2-tailed)	.000	
	N	406	406

** Correlation is significant at the 0.01 level (2-tailed).

The correlation coefficient between the two variables "*Time to Accelerate from 0 to 60 mph (sec)*" and "*Vehicle Weight (lbs.)*" is –0.415, which appears to be weak and negative.



The points are scattered in the above scatter plot. They reveal a weak and negative relationship between the variables. No visible outliers.

SPSS code: Enter data or data from a file → Analyze → Correlate → Bivariate → Move both variables into 'Variables' Box, click 'Pearson', then click OK.

Graphs -> Legacy Dialogs -> Scatter/Dot- > Simple Scatter- > Define -> Move 'Vehicle Weight (lbs.)'(independent variable) into x-axis box and 'Time to Accelerate from 0 to 60 mph (sec)' (dependent variable) into y-axis box → specify title → click OK.

Enter data or read data from a file → Analyze → Regrssion → Linear→ Move 'Vehicle Weight (lbs.)' into the Independent(s) box and 'Time to Accelerate from 0 to 60 mph (sec)' into the Dependent box. Click on 'Statistics', check 'Estimates' and 'Model fit' → Continue → click 'Plots' → move 'ZPRED' in the Y box and 'ZRESID' in the X box → Continue → click OK.

Regression

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	Vehicle Weight (lbs.)(a)	.	Enter

a All requested variables entered.
b Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

Note: "Vehicle weight (lbs.)" is the independent variable and "Time to accelerate from 0 to 60 mph (sec)" is the dependent variable.

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.415(a)	.172	.170	2.569

a Predictors: (Constant), Vehicle Weight (lbs.)
b Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

1. What is the value of R^2? Interpret it.

The value of R^2 is only 0.172, very small. Only 17% of the variation in "*Time to Accelerate from 0 to 60 mph (sec)*" can be explained by this regression model with "*Vehicle Weight (lbs.)*".

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	19.588	.464		42.214	.000
	Vehicle Weight (lbs.)	-.001	.000	-.415	-9.174	.000

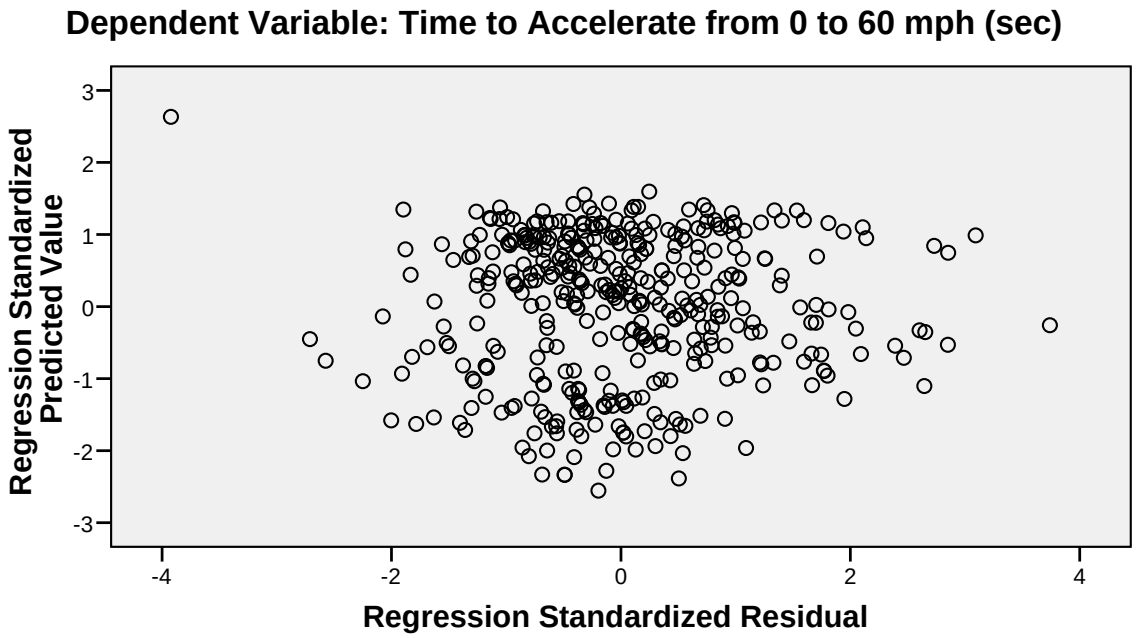
a Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

2. Write down the regression equation from the above table.

SPSS writes intercept as Constant and slope of the equation is identified by the independent variable name. So intercept = 19.588 and slope = –0.001. Therefore the regression equation is given by,

Y-hat = 19.588 –0.001*X*, where *Y-hat*: Predicted Time to Accelerate from 0 to 60 mph (sec) and
X: Vehicle Weight (lbs.).

Scatterplot



This is a scatter plot of the standardized predicted values versus the standardized residuals. The plot reveals a random pattern, except for one point (-4, 3) which could be an outlier. Random pattern in residual plot indicates a linear relationship between the dependent (*Time to Accelerate from 0 to 60 mph (sec)*) and independent (*Vehicle Weight (lbs.)*) variables. The random pattern justifies performing regression analysis for this data.