Regression Techniques

Feb 02, 2017

Outline

Predictive Analytics

- Predicts the probability of occurrence of a future event.
  - Which product the customer is likely to buy in his next purchase? (recommender system).
  - Which customer is likely to default in his/her loan payment? (credit risk).
  - Who is likely to cancel the product that was ordered through e-commerce portal ?
  - How to identify the most profitable customer?
  - What percentage of loans are likely to result in a loss?

Regression Techniques

- Regression is a tool for finding existence of an association relationship between a dependent variable (Y) and one or more independent variables ($X_1, X_2, ..., X_n$) in a study.
- The relationship can be linear or non-linear.

Mathematical Vs Statistical Relationship

1. Mathematical relationship is an exact relationship.
   - $Y = \beta_0 + \beta_1 X$
2. Statistical relationship is not an exact relationship.
   - $Y = \beta_0 + \beta_1 X + \epsilon$

Nomenclature in Regression

- A dependent variable (response variable) "measures an outcome of a study (also called outcome variable)".
- An independent variable (explanatory variable) "explains changes in a response variable".

Table 1: Nomenclature

Dependent Variable	**Independent Variable**
Explained Variable	Explanatory variable
Regressand	Regressor
Predictand	Predictor
Controlled Variable	Control Variable
Target Variable	Stimulus Variable
Response Variable	

Types of Regression

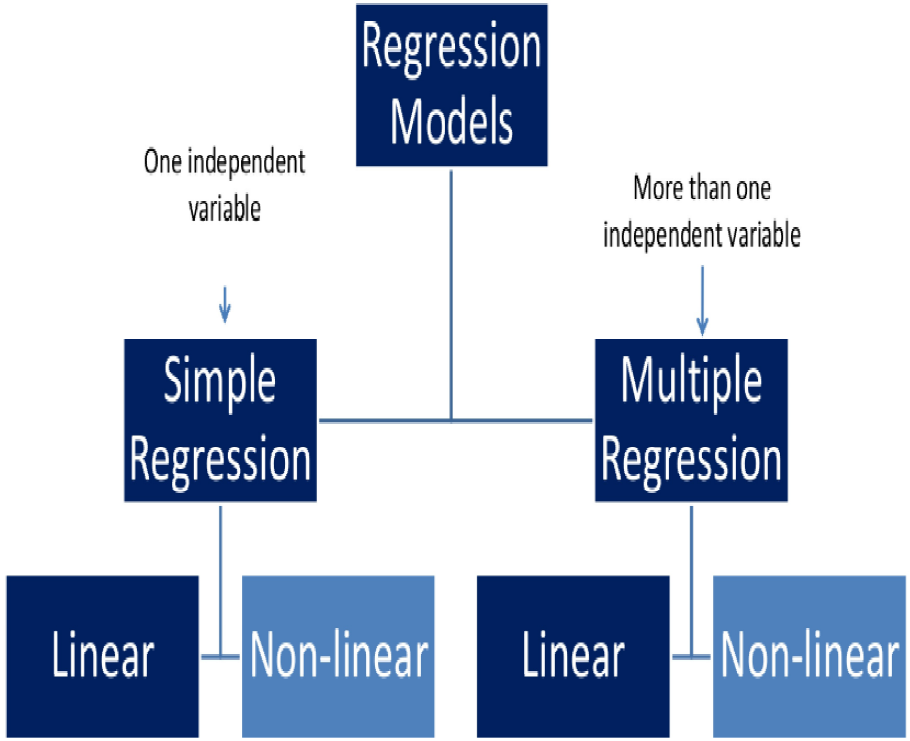


Figure 1: Regrssion types

Types of Regression ...Continued...

1. Simple linear regression
   - $Y = \beta_0 + \beta_1 X + \epsilon$
2. Multiple linear regression
   - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$
3. Nonlinear regression
   - $Y = \beta_0 + \frac{1}{\beta_1 + \beta_2 X_1} + X_2^{\beta_2} + \epsilon$

Regression Model Development

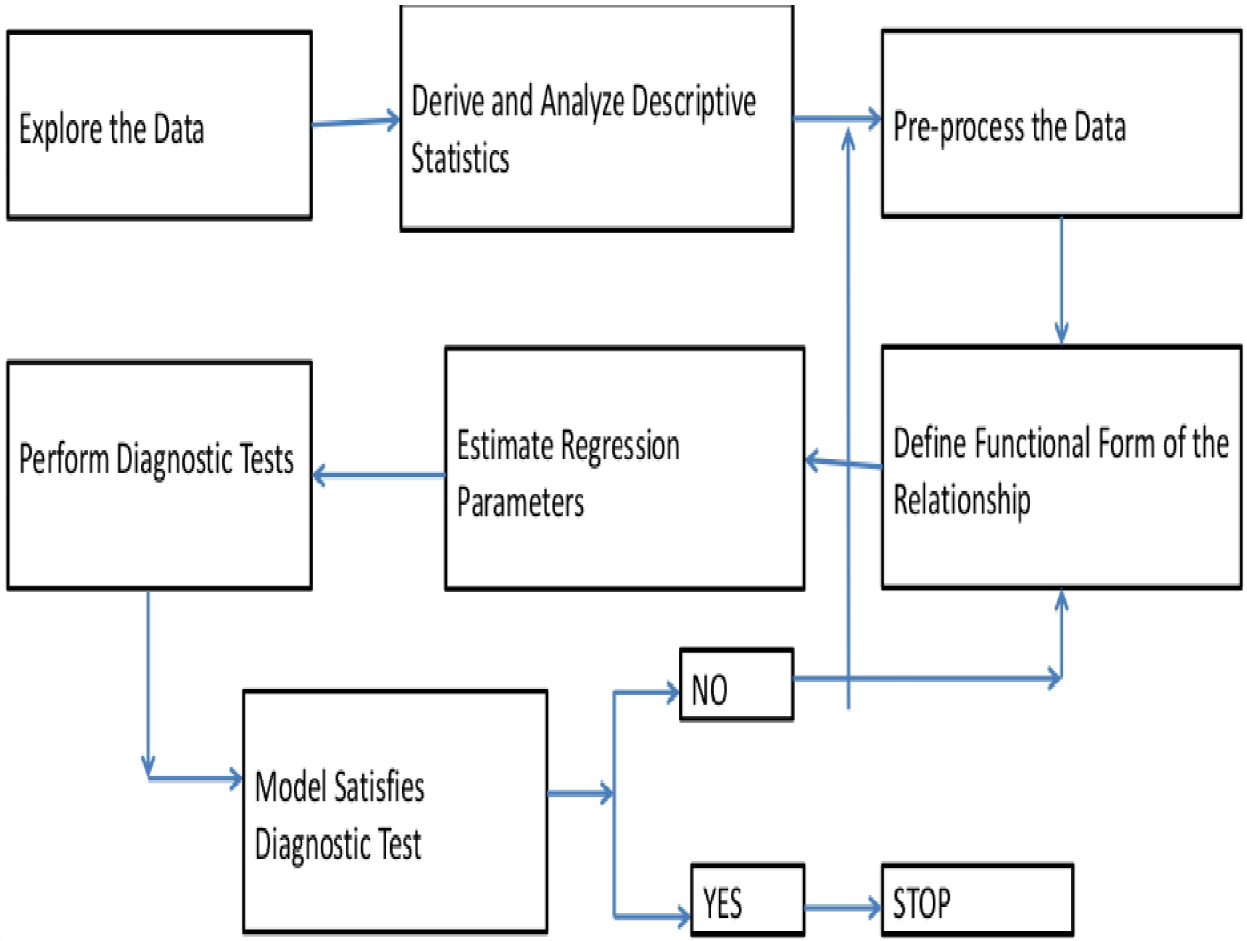


Figure 2: Regression Model Development

Functional Form

- Identify the explanatory variable.
- Specify the nature of relationship between dependent variable and explanatory variables.

Relationship between variables is a linear function:

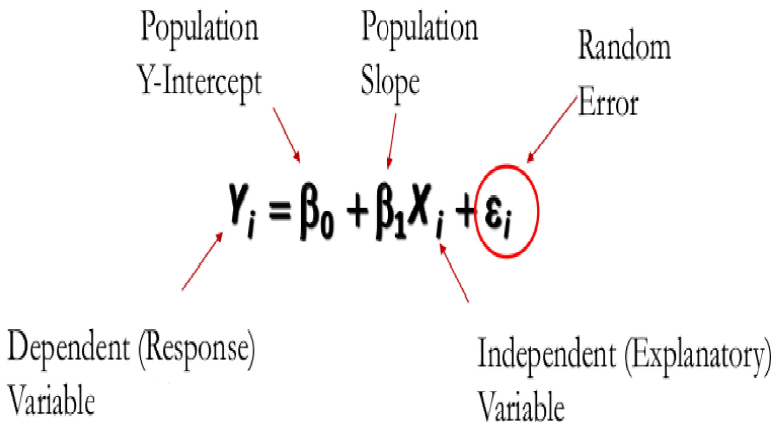


Figure 3: Linear function

Regression OLS estimation

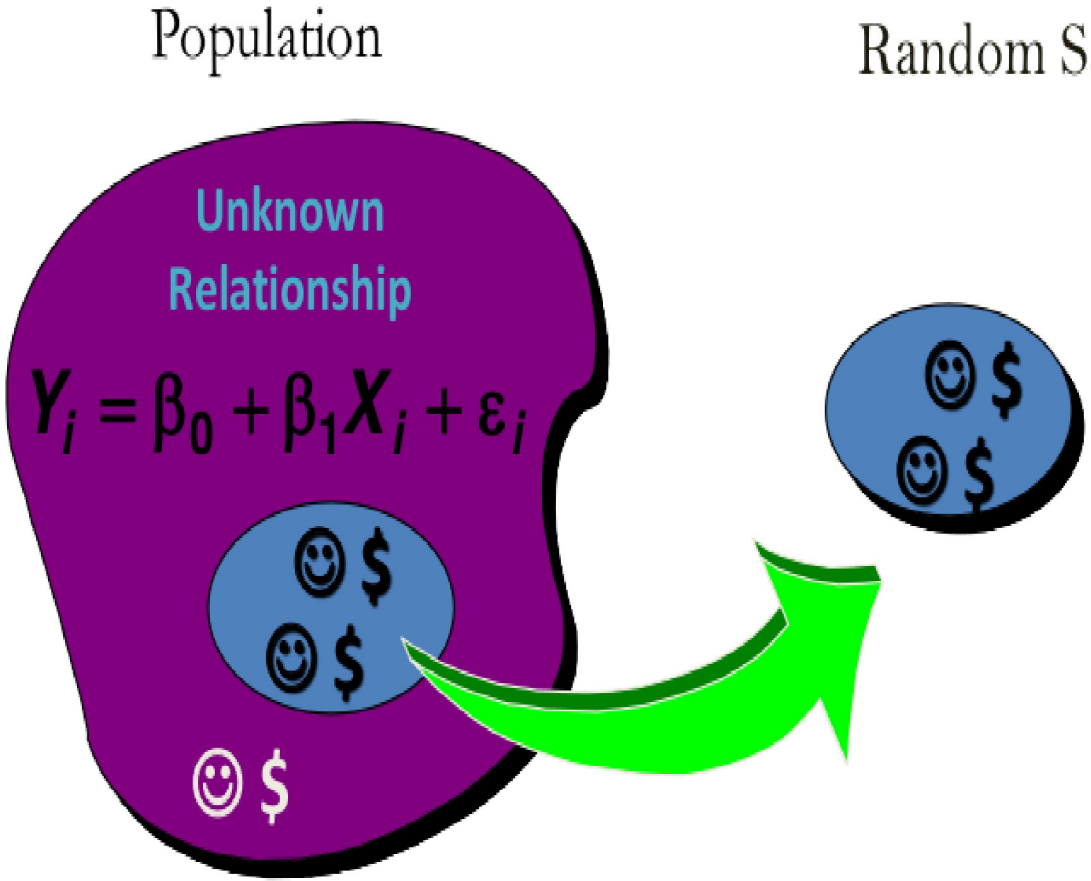
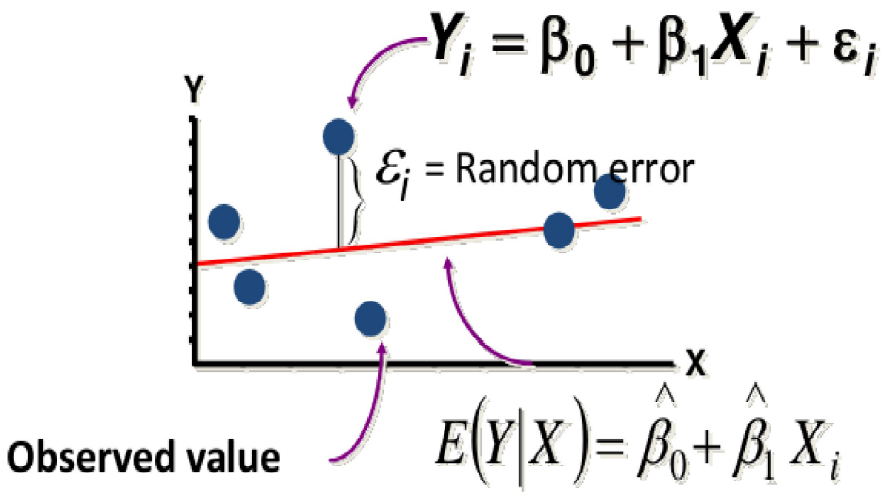
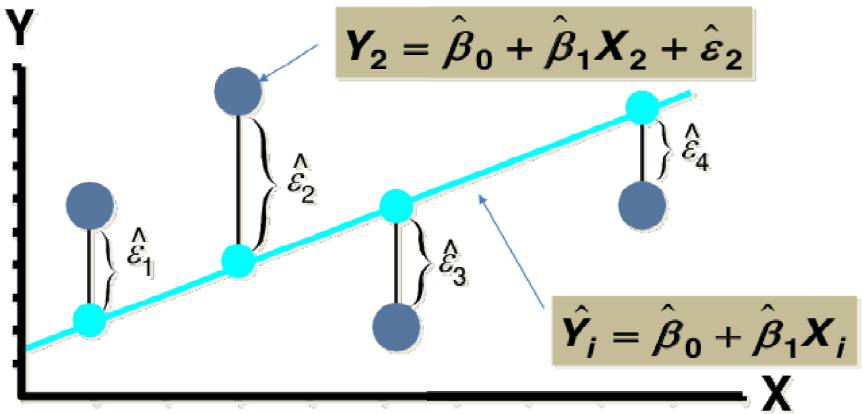


Figure 4: Parameter Estimation

OLS



$$\text{LS minimizes } \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + ... + \hat{\varepsilon}_n^2$$



Estimation of Parameters in Regression

- The least squares function is given by:

$$SSE = \Sigma_{i=1}^{n} \epsilon^2 = \Sigma_{i=1}^{n} (y_i - \beta_0 - \Sigma_{j=1}^{k} \beta_j x_{ij})^2$$

- The least squares estimates must satisfy:

$$\frac{\partial SSE}{\partial \beta_0} = -2\Sigma_{i=1}^{n} (y_i - \hat{\beta}_0 - \Sigma_{j=1}^{k} \hat{\beta}_j x_{ij}) = 0$$

Coefficient Equations

- Prediction Equation:
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Sample Slope:
  - $\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$
  - $\hat{\beta}_1 = \frac{\Sigma_i x_i y_i - n\bar{x}\bar{y}}{\Sigma_i x_i^2 - n(\bar{x})^2}$
- Sample Y-intercept:
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Interpretation of Regression Coefficients

- The interpretation depends on the functional form of the relationship between the response and the explanatory variables.
- The intercept, β_0 , is the mean value of the dependent variable Y, when the independent variable $X = 0$.
- The slope, β_1, is the change in the value of the dependent variable, Y, for unit change in the independent variable X.

Simple Linear Regression

Table 2: Linear Regression

Variable x and y has Linear relationship	Assumption of the world
$y = \beta_0 + \beta_1 x + \epsilon$, Minimize SSE	Fitting a model
Is x really related to y? Is β_1 statistically significant?	Validating the model
Predict y for a given x.	Using a model

Coefficient of determination

- The coefficient of determination (R^2) is a measure of how well the regression line fits the data.
- The value of R^2 lies between 0 and 1 and is the percentage of variation explained by the regression model.
- R^2 is a rough indicator of the worth of the regression model.
- R^2 is the square of the correlation coefficient $r(R^2 = r^2)$.

Variation in Y

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

Variation in Y_i = Systemic Variation + Random Variation

or

Variation in Y_i = Explained Variation + Unexplained Variation

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

Total variation | Explained variation | Unexplained variation

Variation in Y ... Continued...

- TOTAL SUM OF SQUARES (SST):
  - $SST = \Sigma(Y_i - \bar{Y})^2$
  - How much error is there in predicting Y without the knowledge of X?
- SUM OF SQUARES ERROR (SSE):
  - $SSE = \Sigma(Y_i - \hat{Y})^2$
  - How much error is there in predicting Y with the knowledge of X?
- SUM OF SQUARES REGRESSION (SSR):
  - $SSR = \Sigma(\hat{Y}_i - \bar{Y})^2)$
  - Amount of variation explained by the model
- $SST = SSR + SSE$

Coefficient of determination

- Coefficient of determination is the ratio sum of squares due to regression to the total sum of squares.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Example

	Driving Experience (years)	Monthly Auto Insurance Premium
	x	y
1	5	64
2	2	87
3	12	50
4	9	71
5	15	44
6	6	56
7	25	42
8	16	60

Example

	x	y	x^2	xy	$\hat{Y}$	SST	SSE	SSR	SS_x
1	5	64	25	320	68.922	22.5625	24.230	93.555	39.0625
2	2	87	4	174	73.565	770.062	180.494	204.924	85.5625
3	12	50	144	600	58.089	85.562	65.436	1.347	0.5625
4	9	71	81	639	62.732	138.062	68.358	12.124	5.0625
5	15	44	225	660	53.446	232.562	89.237	33.680	14.0625
6	6	56	36	336	67.374	10.5625	129.386	66.012	27.5625
7	25	42	625	1050	37.970	297.562	16.235	452.810	189.0625
8	16	60	256	960	51.898	0.5625	65.626	54.037	22.5625
SUM	90	474	1396	4739		1557.5	639.006	918.493	383.5
Aver-age	11.25	59.25							

β_1	-1.547	R^2	0.589
β_0	76.660	**Standard deviation of error**	10.319
Std deviation	$S(\beta_1)$	0.526	
	S(β_0)	6.961	

$$\hat{y} = 76.660 - 1.547x$$

Example

	Age	**Cholesterol Level**
	x	y
1	58	189
2	69	235
3	43	193
4	39	177
5	63	154
6	52	191
7	47	213
8	31	165
9	74	198
10	36	181

Model validation

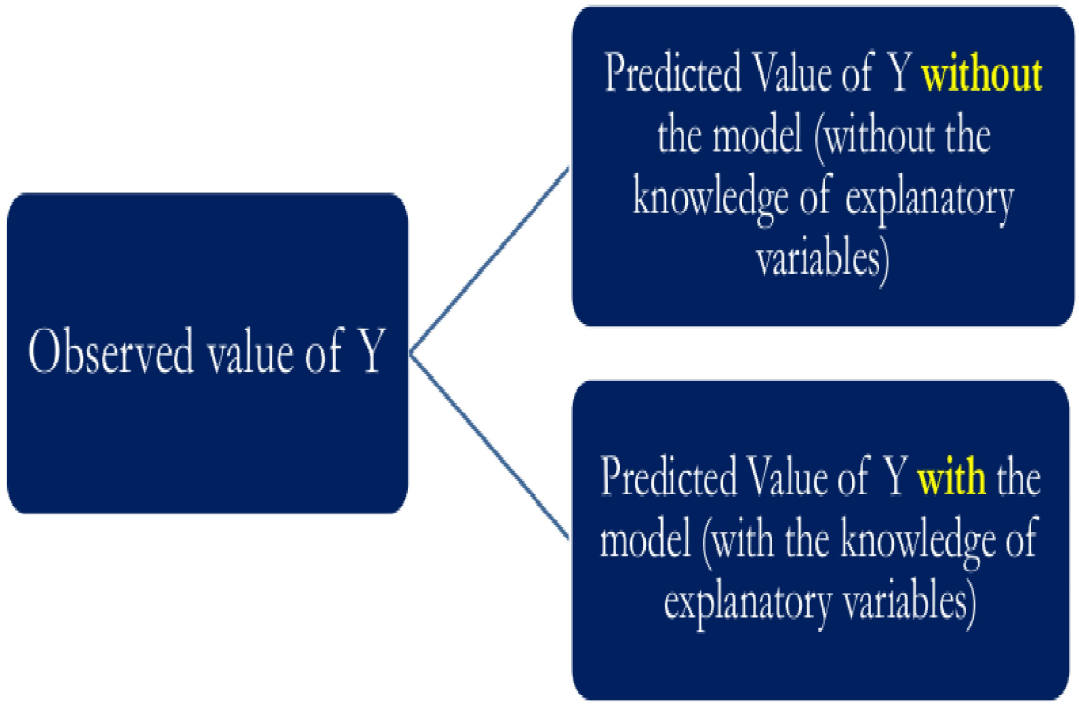


Figure 7: Model validation

- t-test to validate relationship dependent and individual independent variable.

Standard Error of Estimate

- Standard error is the estimate of the standard deviation of the regression errors.
- Standard error of estimate, S_e , measures the variability or scatter of the observed values around the regression line.

$$S_e = \sqrt{\frac{\Sigma(Y_i - \bar{Y}_i)^2}{n-2}}$$

- A smaller standard error of estimate indicates better fit.
- The larger the standard error of estimate, the greater the scattering of points around the regression line.
- If $S_e = 0$, then we can expect a perfect fit.

Standard Error of Estimate for Regression Coefficients

- Standard error of estimate for regression coefficient measures the amount of sampling error in a regression coefficient.
- Standard error of β_0 and β_1
  - $S(\beta_0) = \frac{S_e * \sqrt{\Sigma(x^2)}}{\sqrt{nSS_x}}$
  - $S(\beta_1) = \frac{S_e}{\sqrt{SS_x}}$
  - $SS_x = \Sigma_i (X_i - \bar{X})^2$

T-test

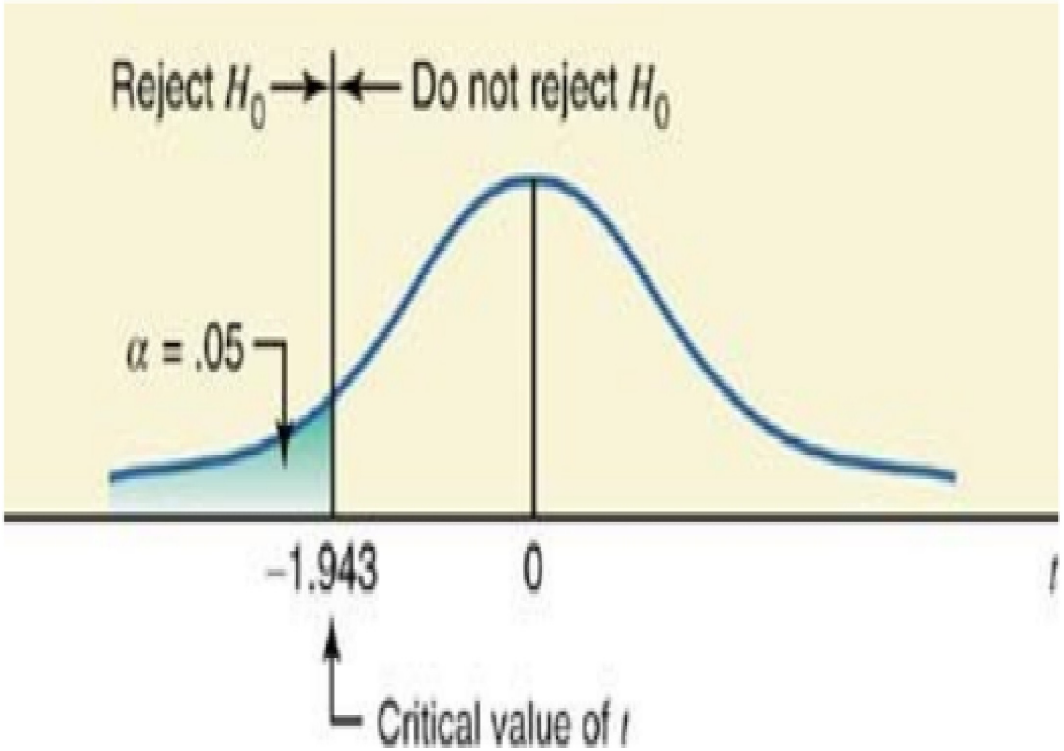
- 90% confidence level, the area in each tail of the t distribution is $\alpha/2 = (1 - 0.90)/2 = 0.05$
- Degree of freedom $df = 8 - 2 = 6$
- From the t distribution table, the t value for .05 area in the right tail of the t distribution and 6 df is 1.943.
- The 90% confidence interval for β_1 is
  - $b \pm ts_{\beta_1}$
  - $-1.5476+1.943(0.5270)$ & -1.5476-$1.943(0.5270)$
  - $-2.57 to -0.52$
- Thus, we can state with 90% confidence that β_1 lies in the interval $-2.57 to -0.52$ That is, on average, the monthly auto insurance premium of a driver decreases by an amount between 0.52 and 2.57 for every extra year of driving experience.

Hypothesis

State the null and alternative hypotheses:

- $H_0 : \beta_1 = 0 (\beta_1 \ is \ not \ negative)$
- $H_0 : \beta_1 < 0 (\beta_1 \ is \ negative)$
- Rejection and non rejection region: From the t distribution table, the critical value of t for .05 area in the left tail of the t distribution and 6 df is 1.943.

Variation in Y



Calculate the value of the test statistic

$$t = \frac{\beta_1 - B}{S_{\beta_1}}$$

$$t = \frac{-1.5476 - 0}{0.5270}$$

$$t = -2.937$$

Decision: Test statistic

- The value of the test statistic $t = -2.937$ falls in the rejection region. Hence, we reject the null hypothesis and conclude that β_1 is negative.
- That is, the monthly auto insurance premium decreases with an increase in years of driving experience.

Decision: P value

- Find the range for the p-value from the t distribution table and make a decision by comparing that p-value with the significance level.
- Here df = 6 and the observed value of t is -2.937.
- From the t distribution table in the row of df = 6, 2.937 is between 2.447 and 3.143.
- The corresponding areas in the right tail of the t distribution are .025 and .01.
- But our test is left-tailed and the observed value of t is negative. Thus, t = -2.937 lies between -2.447 and -3.143.

Decision: P value ... Continued ...

- The corresponding areas in the left tail of the t distribution are .025 and .01. Therefore the range of the p-value is $.01 < pvalue < .025$
- Thus, we can state that for any α equal to or greater than .025, we will reject the null hypothesis.
- Here, $\alpha = .05$, which is greater than the upper limit of the p value of .025.
- As a result, we reject the null hypothesis. If we use technology to find this p value, we will obtain a p value of .013. Then we can reject the null hypothesis for any $\alpha > .013$.

THANK YOU