data analytics

UNIT-2

Regression Analysis

feedback/corrections: vibha@pesu.pes.edu

VIBHA MASTI

CORRELATION ANALYSIS

- Correlation ≠ causation

- Strength and direction of a relationship between two random variables

- Choose variables for model building

- Correlation types
  - Between continuous (interval, ratio) variables
  - Between ordinal variables
  - Between continuous RV and dichotomous (binary) RV
  - Between two binary variables

1. Pearson's Correlation Coefficient

- Pearson Product Moment Correlation

- Strength and direction of linear relationship between two cont. RVs

$$\rho = r = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \qquad \text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

TABLE 8.1 Data on age and average call duration (in seconds)

Age	14	15	18	19	20	24	25	27	29	30
Call Duration	540	544	567	548	550	520	512	516	511	511
Age	33	36	38	39	40	41	42	43	45	48
Call Duration	490	487	472	460	455	463	440	422	411	397

T1

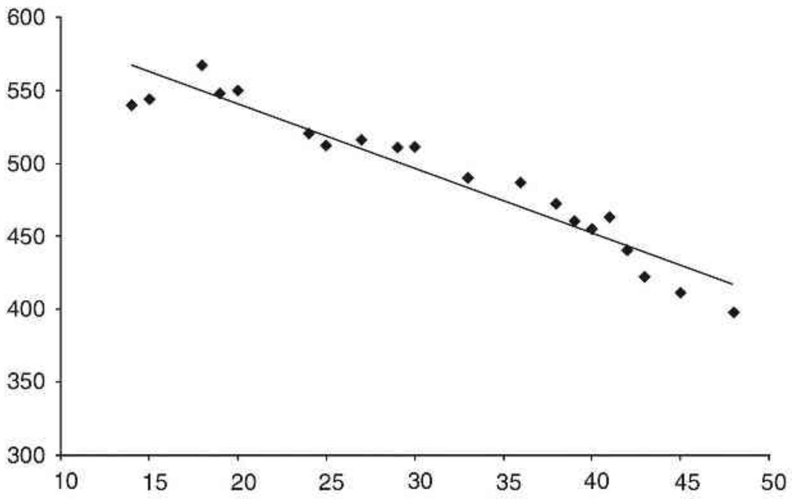


FIGURE 8.1 Association relationship between age and average call duration.

T1

- Values must be normalised (using z statistics)

$$Z_x = \frac{X - \bar{X}}{\sigma_x} \qquad Z_y = \frac{Y - \bar{Y}}{\sigma_y}$$

- Pearson's correlation coefficient given by

$$r = \frac{\sum_{i=1}^{n} z_{xi} \, z_{yi}}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n \, \sigma_x \, \sigma_y}$$

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \, s_x \, s_y}$$

- Correlation varies from -1 to +1
- Simplification of r

$$r = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \times \sqrt{n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

- If r is +ve, positive correlation and if r is -ve, negative correlation
- If r is correlation coefficient for X & Y. Then the correlation coefficient for $Z_1 = aX + b$ and $Z_2 = cY + d$ is r if the signs of a & c are same and -r if the signs are different
- Coefficient of Determination $R^2 = r^2$

Q: The average share prices of two companies over the past 12 months are shown in the table. Calculate the Pearson correlation coefficient.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

$$\left.\begin{aligned} \bar{x} &= 292.972 \\ \bar{y} &= 229.829 \\ \Sigma xy &= 811242.980 \\ \Sigma x^2 &= 1035905.663 \\ \Sigma y^2 &= 637209.3349 \end{aligned}\right\}$$ solved using fx-991 EX calculator

$$r = \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{\Sigma x_i^2 - n\bar{x}^2}\sqrt{\Sigma y_i^2 - n\bar{y}^2}}$$

$r = 0.728$ } can calculate directly with calc

∴ X & Y are positively, strongly correlated

Spurious Correlation

- Due to hidden variables
- C → A & C → B, does A → B
- Eg: stork population & human birth rate — hidden variable: available nesting area
- Eg: doctors & deaths: Young, 2001
- Eg: Divorce rate in Maine and per capita consumption of margarine: tylervigen.com
- Correlation ⇏ relation

Hypothesis Test for Correlation Coefficient

H_0:	$\rho = 0$ (there is no correlation between two random variables)
H_A:	$\rho \neq 0$ (there is a correlation between two random variables)

- Sampling distribution of correlation coefficient r follows t distribution with n-2 degrees of freedom (df)
- 2 df lost because we estimate two mean values from the data

- The mean of the sampling distribution is ρ (population correlation coefficient)

- Standard deviation of the sampling distribution is

$$\sqrt{\frac{1-r^2}{n-2}}$$

- The t-statistic for the null hypothesis $U = t_{\alpha/2, n-2} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$

- When $\rho = 0$

$$U = t_{\alpha/2, n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Q: The average share prices of two companies over the past 12 months are shown in the table. Conduct the following two hypothesis tests at $\alpha = 0.05$:
(a) The correlation between share prices of two companies is zero.
(b) The correlation between share prices of two companies is at least 0.5.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

(a) $H_0: \rho = 0$ $\alpha = 0.05$
$H_a: \rho \neq 0$

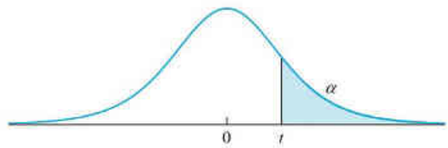
$r = 0.7279$

$$U = t_{\alpha/2, 10} = \frac{0.7279}{\sqrt{\frac{1-0.7279^2}{10}}} = 3.357$$

(two-tailed test)

$\alpha/2$ is between 0.005 and 0.001
α is between 0.01 and 0.002

TABLE A.3 Upper percentage points for the Student's t distribution



ν \ α	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587

∴ We can reject H_0 and accept H_a ⇒ the correlation in share prices is not 0

(b) $H_0: \rho \geq 0.5$
$H_a: \rho < 0.5$

$$U = t_{\alpha, 10} = \frac{0.7279 - 0.5}{\sqrt{\frac{1 - 0.7279^2}{10}}} = 1.051$$

Right-tailed test

α is between 0.10 and 0.25

∴ we cannot reject Ho as the p-value > 0.05

2. Spearman Rank Correlation

- Ordinal variables (ρ_s - population, r_s - sample)

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

D_i = difference in the rank of case i $(x_i - y_i)$

- Sampling distribution of r_s follows t-distribution with mean ρ_s and SD with n-2 degrees of freedom

$$S = \sqrt{\frac{1-r_s^2}{n-2}}$$

Q: Ranking of 12 countries under corruption and Gini Index (wealth discrimination) are shown in Table. Calculate the Spearman correlation and test the hypothesis that the correlation is at least 0.2 at $\alpha = 0.02$.

Countries	1	2	3	4	5	6	7	8	9	10	11	12
Corruption	1	4	12	2	5	8	11	7	10	3	6	9
Gini Index	2	3	9	5	4	6	10	7	8	1	11	12

$$r_s = 1 - \frac{6(1+1+9+9+1+4+1+0+4+4+25+9)}{12(143)}$$

$$= 1 - \frac{34}{143} = 0.7622$$

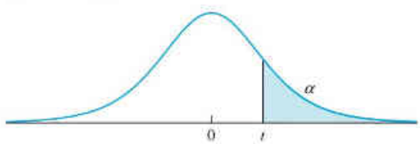
Country	Corruption Rank (X_i)	Gini Index (Y_i)	$D = X_i - Y_i$	D^2
1	1	2	−1	1
2	4	3	1	1
3	12	9	3	9
4	2	5	−3	9
5	5	4	1	1
6	8	6	2	4
7	11	10	1	1
8	7	7	0	0
9	10	8	2	4
10	3	1	2	4
11	6	11	−5	25
12	9	12	−3	9
			$\sum_{i=1}^{12} D_i^2$	68

$H_0: \rho < 0.2$ $\quad\quad \alpha = 0.02$

$H_a: \rho \geq 0.2$

$$U = t_{0.02,10} = \frac{0.7622 - 0.2}{\sqrt{\frac{1 - 0.7622^2}{10}}} = 2.746$$

TABLE A.3 Upper percentage points for the Student's t distribution



ν	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	[illegible]

p-value between 0.025 and 0.01

Critical point found (Excel) gives more info; critical point for $\alpha = 0.02$ is 2.35

$\therefore$ can reject H_0 and accept H_a

3. Point Bi-Serial Correlation

- Correlation between a continuous RV and a dichotomous RV
- Group x instances into 2 groups: where $y = 0$ and where $y = 1$
- n_0 = no. of instances with $x = 0$ and n_1 = no. of instances with $x = 1$
- Pearson's Point Bi-Serial correlation

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_x} \sqrt{\frac{n_0 n_1}{n(n-1)}}$$

Q: Ms Sandra Ruth, data scientist at Airmobile, is interested in finding the correlation between the average call duration and gender. The table provides the average call duration (measured in seconds) and gender of 30 customers of Airmobile. In the table, male is coded as 0 and Female is coded as 1. Calculate the point bi-serial correlation.

Gender	1	1	0	0	0	1	0	1	1	0
Call Duration	448	335	210	382	407	231	359	287	288	347
Gender	1	1	1	1	1	0	0	1	0	0
Call Duration	408	382	303	201	447	439	383	277	279	213
Gender	1	1	0	1	1	0	1	0	1	0
Call Duration	383	355	362	401	331	421	367	437	326	351

$\bar{x}_0 = 345.07$ $\quad \bar{x}_1 = 339.412$ $\quad \bar{x} = 345.33$

$S_x = 71.7189$ $\quad n_0 = 13$ $\quad n_1 = 17$

$$r_b = -0.0960$$

∴ Very low negative correlation

4. Phi Coefficient

- Both RVs are binary
- Contingency table

Y

X

	$\overline{N_y}$	N_y	
$\overline{N_x}$	N_{00}	N_{01}	N_{0x}
N_x	N_{10}	N_{11}	N_{1x}
	N_{y0}	N_{y1}	N

N_{00} = Number of cases in the sample such that $X = 0$ and $Y = 0$
N_{01} = Number of cases in the sample such that $X = 0$ and $Y = 1$
N_{10} = Number of cases in the sample such that $X = 1$ and $Y = 0$
N_{11} = Number of cases in the sample such that $X = 1$ and $Y = 1$
N_{X0} = Number of cases in the sample such that $X = 0$
N_{X1} = Number of cases in the sample such that $X = 1$
N_{Y0} = Number of cases in the sample such that $Y = 0$
N_{Y1} = Number of cases in the sample such that $Y = 1$

$$\phi = \frac{N_{11} N_{00} - N_{10} N_{01}}{\sqrt{N_{X0} N_{X1} N_{Y0} N_{Y1}}}$$

Q: Joy Finance (JF) is a company that provides gold loans (in which gold is used as guarantee against the loan). Mr Georgekutty, Managing Director of JF, collected data to understand the relationship between loan default status (variable Y) and the marital status of the customer (variable X). Data is collected on past 40 loans and is shown in Table 8.8. Calculate the Phi- coefficient. In Table , Y = 0 implies non-defaulter, Y = 1 is defaulter, X = 0 is single, and X = 1 is married.

TABLE 8.8 Marital status (0 = Single, 1 = Married) versus loan status (0 = No default, 1 = Default)

X	1	0	1	0	0	0	0	0	1	0
Y	0	1	0	1	0	0	0	1	1	1
X	0	1	1	0	0	1	0	0	0	1
Y	0	1	1	1	0	0	1	1	0	0
X	1	0	0	0	1	1	1	0	0	1
Y	0	0	0	1	0	0	0	0	0	0
X	1	0	0	0	1	0	1	0	1	1
Y	1	0	0	1	1	0	1	1	0	1

Contingency table

	Y			Total
		0	1	
X	0	13	10	23
	1	10	7	17
Total		23	17	40

$$\phi = \frac{N_{11} N_{00} - N_{10} N_{01}}{\sqrt{N_{x0} N_{x1} N_{y0} N_{y1}}} = \frac{7 \times 13 - 10 \times 10}{\sqrt{23 \times 17 \times 23 \times 17}} = -0.0230$$

Weak negative correlation

REGRESSION

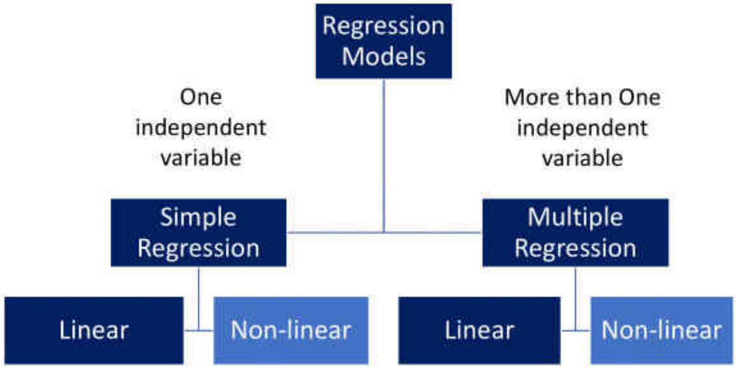
- If there is a significant Linear Correlation between RVs X & Y, one of the 5 can be true
  1. Direct cause & effect relationship
  2. Reverse cause & effect relationship
  3. Maybe due to third variable
  4. Complex interactions of several variables
  5. Coincidental

- Independent variable: explanatory variable
  Dependent variable: response variable

- Regression: supervised learning

- Does not capture causality

Dependent Variable	Independent Variable
Explained Variable	Explanatory variable
Regressand	Regressor
Predictand	Predictor
Endogenous Variable	Exogenous Variable
Controlled Variable	Control Variable
Target Variable	Stimulus Variable
Response Variable	Feature
Outcome Variable	



1. Simple Linear Regression
– two variables

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

2. Multiple Linear Regression
- more than 1 independent variable

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

3. Non-linear Regression

$$Y = \beta_0 + \frac{1}{\beta_1 + \beta_2 X_1} + X_2^{\beta_3} + \varepsilon$$

Linear Regression

- Linear relationship between dependent variables and regression coefficients

$$Y = \beta_1 + \beta_1 X_1 + \beta_2 X_2 X_1 + \beta_3 X_1^2$$

linear regression

- Least squares approach used

1. The regression model is linear in regression parameters.
2. The explanatory variable, X, is assumed to be non-stochastic (i.e., X is deterministic).
3. The conditional expected value of the residuals, $E(\varepsilon_i|X_i)$, is zero.
4. In case of time series data, residuals are uncorrelated, that is, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
5. The residuals, ε_i, follow a normal distribution.
6. The variance of the residuals, $\mathrm{Var}(\varepsilon_i|X_i)$, is constant for all values of X_i. When the variance of the residuals is constant for different values of X_i, it is called **homoscedasticity**. A non-constant variance of residuals is called **heteroscedasticity**.

Functional Form of Relationship

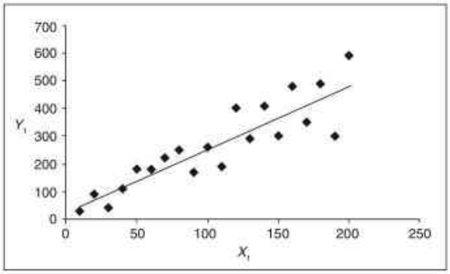


FIGURE 9.3 Linear relationship between X_1 and Y_1.

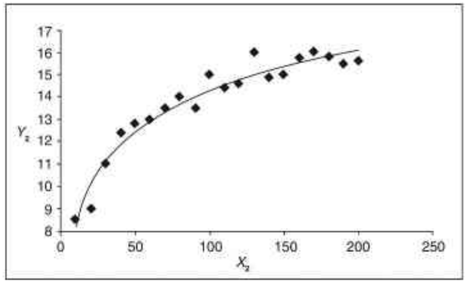


FIGURE 9.4 Log-linear relationship between X_2 and Y_2.

T1

Ordinary Least Squares (OLS) Estimation

- Minimum sum of squared error (SSE)

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Partial derivatives

$$\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n} -2(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$2\left(n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} -2(Y_i - \beta_0 - \beta_1 X_i)X_i = 0$$

$$= -2 \sum_{i=1}^{n} (X_i Y_i - \beta_0 X_i - \beta_1 X_i^2) = 0$$

- Line

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y})}{\sum_{i=1}^{n} (x_i^2 - x_i \bar{x})} = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/#:~:text=The%20Gauss%2DMarkov%20theorem%20states%20that%20if%20your%20linear%20regression,of%20all%20possible%20linear%20estimators

- OLS provides Best Linear Unbiased Estimate (BLUE)

Steps : Framework for SLR Model Development

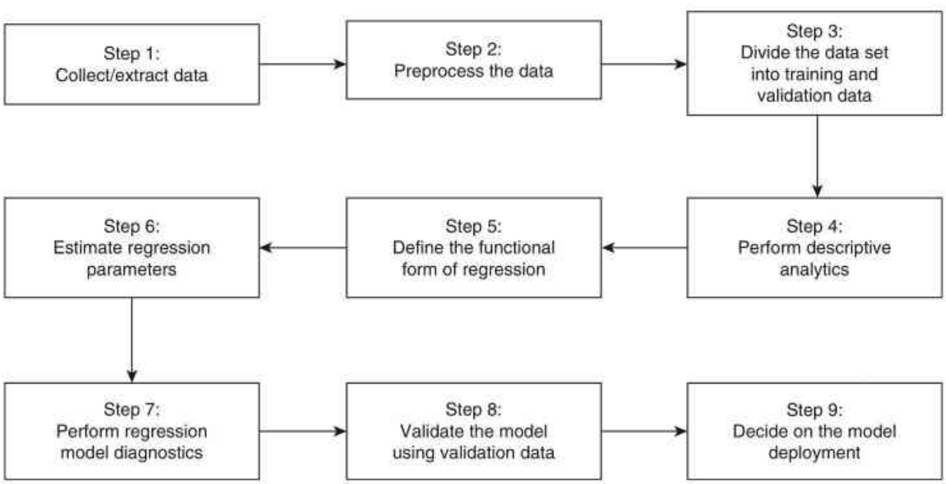


FIGURE 9.2 Framework for SLR model development.

Step 7:

- Regression model dignostics to be performed before applying a model
- Assumptions should not be violated

Step 8:

- Model must be validated against validation dataset
- Prevent overfitting

Q: Table below provides the salary of 50 graduating MBA students of a Business School in 2016 and their corresponding percentage marks in grade 10. Develop a linear regression model by estimating the model parameters.

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
1	62	270000	26	64.6	250000
2	76.33	200000	27	50	180000
3	72	240000	28	74	218000
4	60	250000	29	58	360000
5	61	180000	30	67	150000
6	55	300000	31	75	250000
7	70	260000	32	60	200000
8	68	235000	33	55	300000
9	82.8	425000	34	78	330000
10	59	240000	35	50.08	265000
11	58	250000	36	56	340000
12	60	180000	37	68	177600
13	66	428000	38	52	236000
14	83	450000	39	54	265000
15	68	300000	40	52	200000
16	37.33	240000	41	76	393000
17	79	252000	42	64.8	360000
18	68.4	280000	43	74.4	300000
19	70	231000	44	74.5	250000
20	59	224000	45	73.5	360000
21	63	120000	46	57.58	180000
22	50	260000	47	68	180000
23	69	300000	48	69	270000
24	52	120000	49	66	240000
25	49	120000	50	60.8	300000

Let x = class 10 percent
y = salary

$\hat{\beta}_0 = 61555.3553$

$\hat{\beta}_1 = 3076.1774$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{y} = 61555.3553 + 3076.1774$

Interpretation of SLR

- If functional form is $Y = \hat{\beta}_0 + \hat{\beta}_1 X$

↪ $\hat{\beta}_1 : \frac{\partial Y}{\partial X}$: partial derivative of Y wrt X

↪ $\hat{\beta}_0 : E(Y \mid X=0)$: expected value of Y when $X = 0$

Validation of SLR Model

- Ensure validity & goodness of fit
- Following measures

1. Prediction accuracy
2. Residual analysis
3. Coefficient of determination (R^2)
4. Hypothesis test for regression coefficient (β_1)
5. Analysis of variance
6. Outier analysis

RESIDUAL ANALYSIS

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

- Eg: lemonade stand dataset

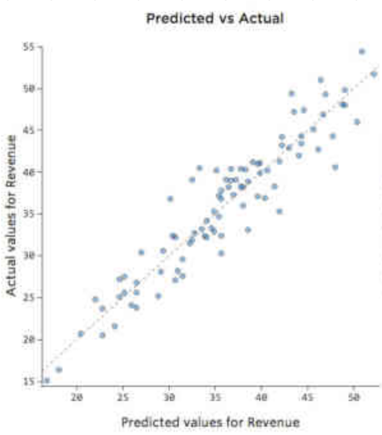
Temperature (Celsius)	Revenue
28.2	$44
21.4	$23
32.9	$43
24.0	$30
etc.	etc.

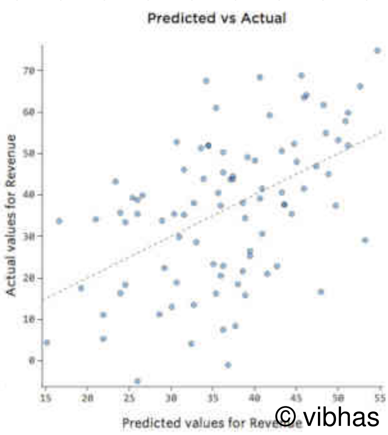
- Regression equation

Revenue = 2.7 × Temperature − 35

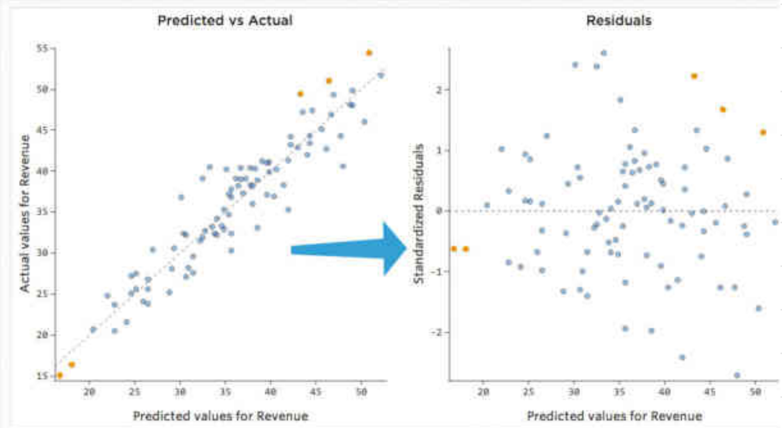
Residual = Observed − Predicted

- Accuracy with observed vs Predicted for 2 diff datasets



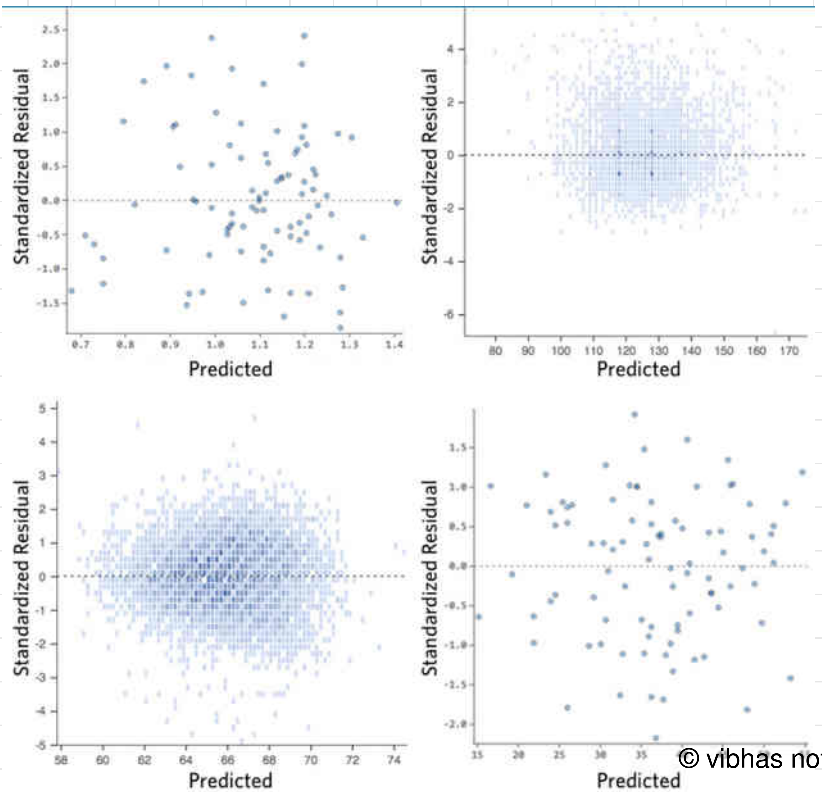


- Residual plot



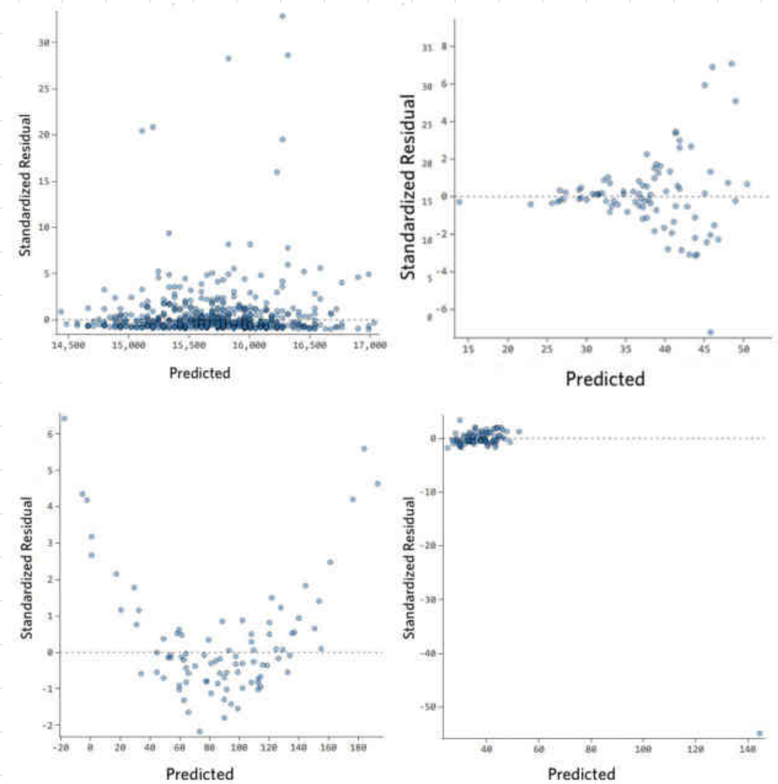
Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.

- For good linear models, the residual plots are evenly distributed with no clear pattern



- As shown above, the four plots
  - Are symmetrically distributed, tending to cluster towards the middle of the plot
  - Are clustered towards lower single digits of the y-axis (close to y=0)
  - Do not show any clear patterns

- Models that could be improved show patterns in the residual plot, indicating that the particular regression chosen is not the best



- As shown above, the four plots
  - Aren't evenly distributed
  - Have an outlier
  - Have a clear shape to them

HOMOSCEDASTICITY

- A good regression model is assumed to be homoscedastic
- In other words, variance of residuals is constant across different values of x
- Variance of residuals is independent of X
- If assumption is not met, the hypothesis tests become unreliable

TI

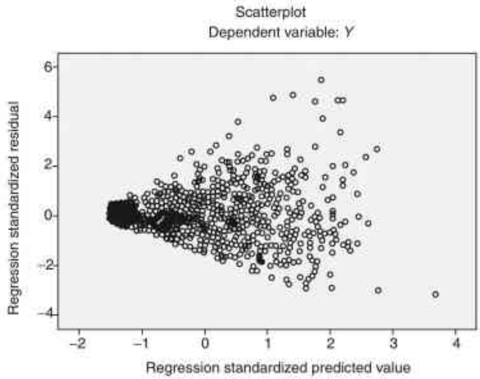


FIGURE 9.7 Funnel shape in the standardized residual plot indicates heteroscedasticity.

- If the residual plot is heteroscedastic, functional form of the regression model used is incorrect (misspecification)

TI

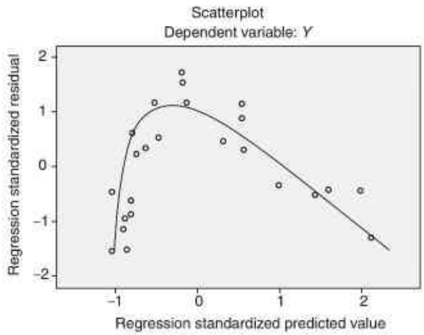


FIGURE 9.8 A pattern (parabola) in the residual plot indicates model misspecification.

Residual Analysis

- Check if assumptions of regression model satisfied

1. Residuals $(Y_i - \hat{Y}_i)$ are normally distributed
2. Variance of residuals is constant — homoscedasticity
3. The functional form of regression is correct
4. There are not too many outliers

1. Normal Distribution of Residuals

- Using P-P (Probability- Probability) plot

- Compares cdf of the two probability distributions with each other

- Compare if residuals of testing variables matches with residuals of normal distribution

TI

(from normal dist with same mean & std dev)

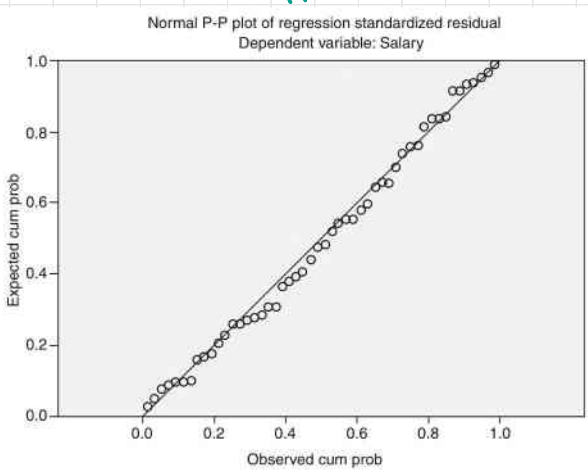


FIGURE 9.5 Residual plot (P-P Plot) of the regression model $Y = \beta_0 + \beta_1$ salary (Example 9.1).

(page 18 of notes)

TI

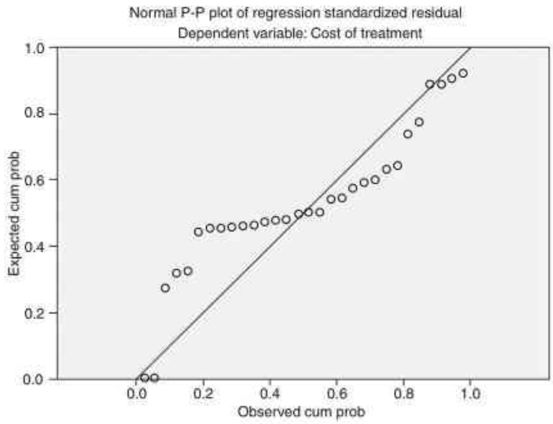


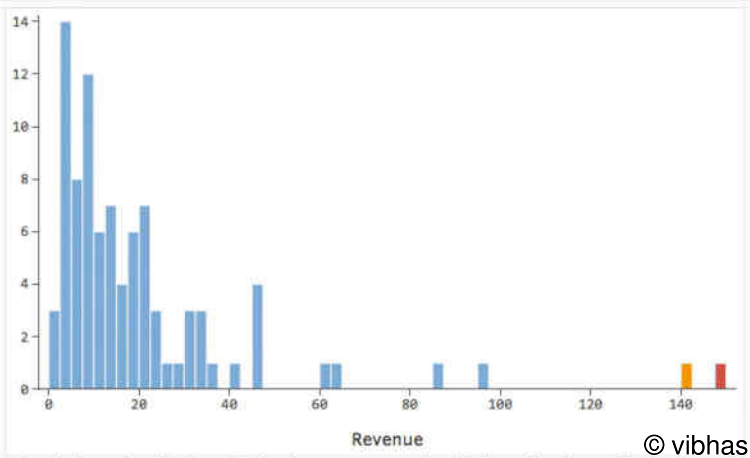
FIGURE 9.6 Residual plot (P - P Plot) of the regression model $Y = \beta_0 + \beta_1$ age (Example 9.2).

2. Variance of Residuals is constant

Fix 1 : transforming the variable

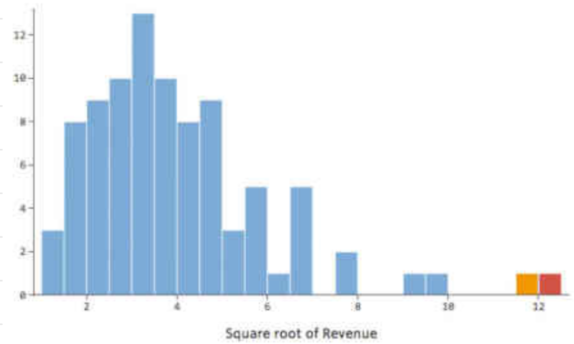
- Common: transform one into log form
- Goal: to get bell-shaped curve

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

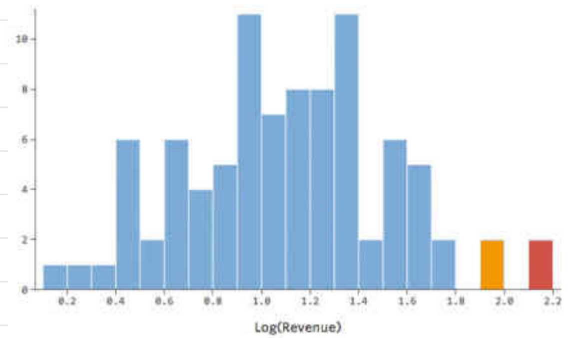


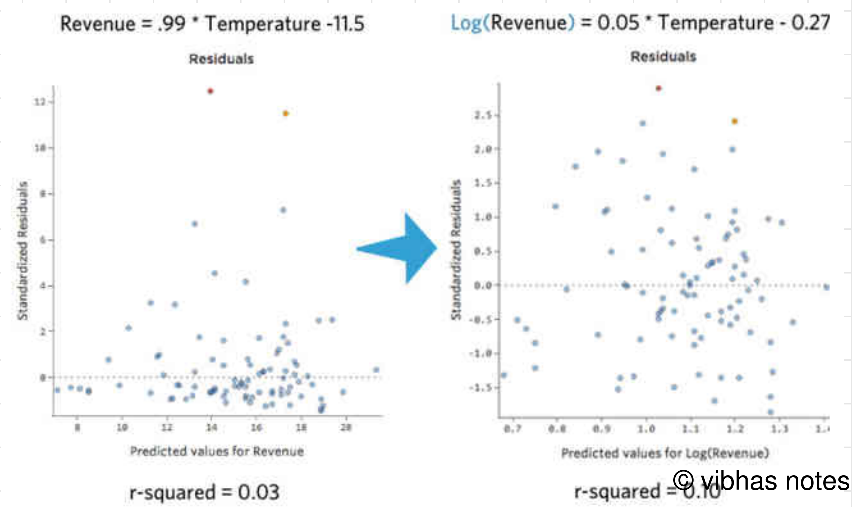
We've colored a couple outlying datapoints red and orange so you can see how their distance from the rest of the data changes after we transform the data.

- Square root transformation



- Log transformation

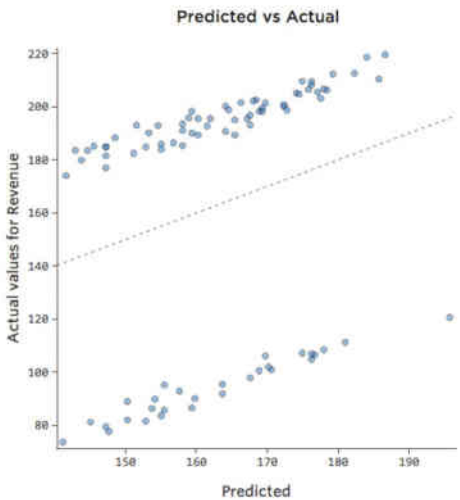




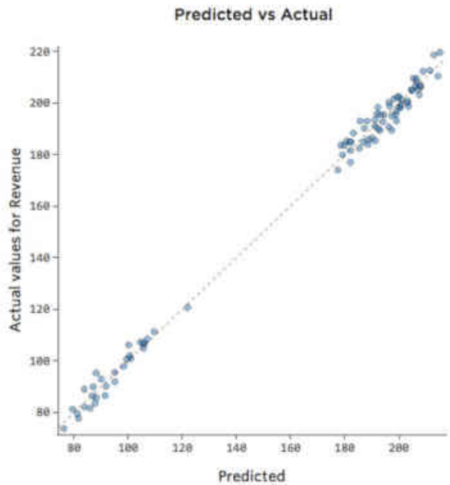
Fix 2: Adding a New Variable

- Eg: lemonade stands are more busy over the weekend

without weekend



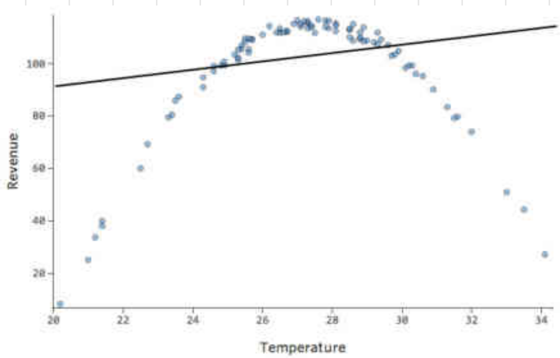
with weekend



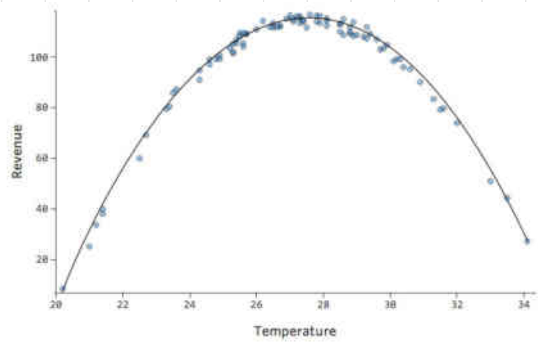
3. Functional form of Regression

- Resort to non-linear model

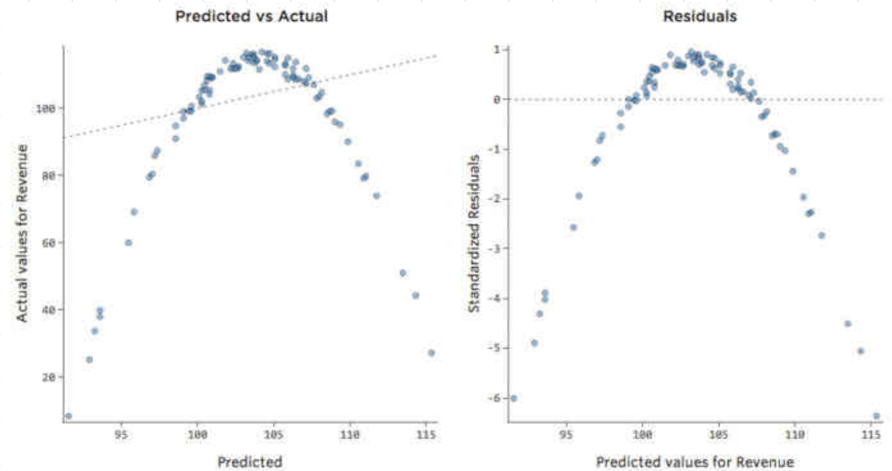
$y = 1.7x + 51$



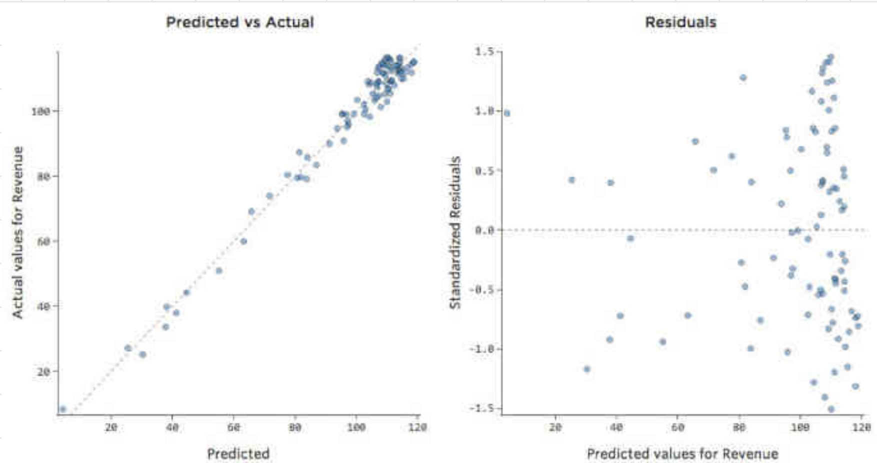
$-2x^2 + 111x - 1408$



- Without x^2



- With x^2



4. Outlier Analysis

- Observations whose values show a large deviation from mean value — outliers

- Can significantly affect values of regression coefficients

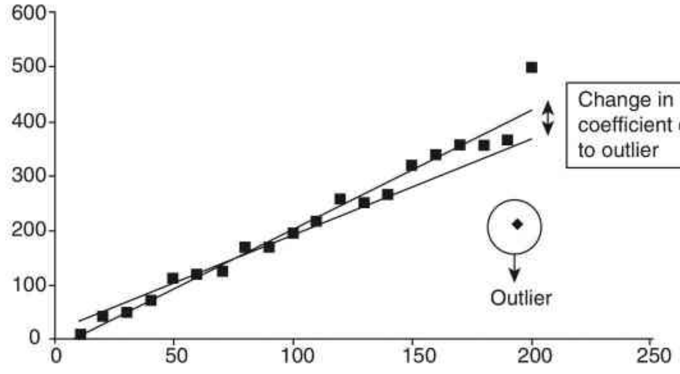
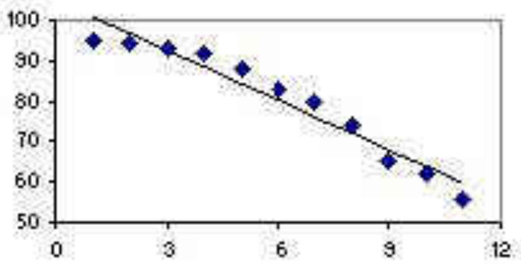


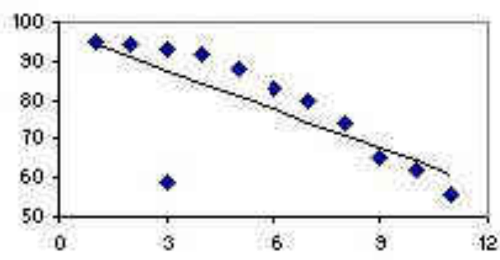
FIGURE 9.9 Influence of outliers on regression coefficients.

Effect of Influential Points



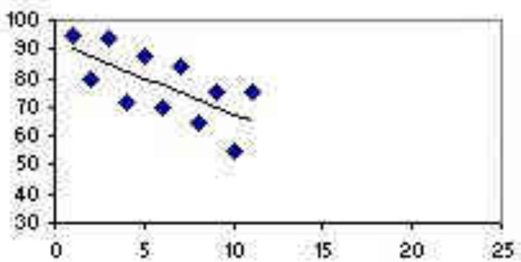
$\hat{y} = 104.78 - 4.10x$

$R^2 = 0.94$



$\hat{y} = 97.51 - 3.32x$

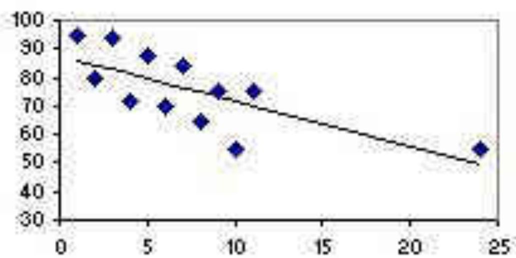
$R^2 = 0.55$



$\hat{y} = 92.54 - 2.5x$

slope $b_0 = -2.5$

$R^2 = 0.46$



$\hat{y} = 87.59 - 1.6x$

slope $b_0 = -1.6$

$R^2 = 0.52$

Regression vs Correlation

- Both: strength of relationship
- Correlation: commutative ; assumes both variables to be random
- Regression : what is the change in Y for a unit change in X
- https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/

Q: You have two measuring systems and you want to see how well they agree with each other. So you measure the same 20 parts with each measuring system

correlation (interchangeable)

Q: You want to predict blood pressure for different doses of a drug

regression (prediction)

Q: A clinical trial has multiple endpoints and you want to know which pair of endpoints has the strongest linear relationship

correlation (scatterplot / correlation matrix)

Q: You want to know how much the response (Y) changes for every one unit increase in (X)

regression (slope)

Coefficient of Determination (R^2)

- SLR : explained variation & unexplained variation

$$\underbrace{Y_i}_{\text{variation in } Y} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{variation in Y explained by the model}} + \underbrace{\varepsilon_i}_{\text{variation in Y not explained by the model}}$$

- Total variation = difference in Y_i and $\bar{Y}$

Variation Type	Measure	Description
Total Variation (SST)	$(Y_i - \bar{Y})^2$	Total variation is the difference between the actual value and the mean value
Variation explained by the model (SSR)	$(\hat{Y}_i - \bar{Y})^2$	Variation explained by the model is the difference between the estimated value of Y_i and the mean value of Y
Variation not explained by model (SSE)	$(Y_i - \hat{Y}_i)^2$	Variation not explained by the model is the difference between the actual value and the predicted value of Y_i (error in prediction)

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SSE}$$

- SST: sum of squares of total variation
- SSR: sum of squares of variation explained by the regression model
- SSE: sum of squares of errors or unexplained variation

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SSR}{SST} = \frac{(\hat{Y}_i - \bar{Y})^2}{(Y_i - \bar{Y})^2}$$

$$SSR = SST - SSE$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{(\hat{Y}_i - Y_i)^2}{(Y_i - \bar{Y})^2}$$

- min value of $r^2 = R^2$ given α can be determined using F-statistic

Spurious Regression

Year	Number of Facebook users in millions (X)	Number of people who died of helium poisoning in UK (Y)
2004	1	2
2005	6	2
2006	12	2
2007	58	2
2008	145	11
2009	360	21
2010	608	31
2011	845	40

The R-square value for regression model between the number of deaths due to helium poisoning in UK and the number of Facebook users is 0.9928. That is, 99.28% variation in the number of deaths due to helium poisoning in UK is explained by the number of Facebook users.

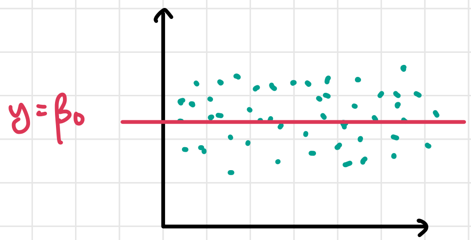
Hypothesis Test for Regression Coefficient (t-test)

- Regression coefficient = β_1 (slope)

$$\beta_1 = \frac{\sum_{i=1}^{n} k_i Y_i}{\sum_{i=1}^{n} k_i^2}$$

$k_i = (X_i - \bar{X})$
can be treated as constant (non-stochastic)

- If $\beta_1 = 0$, no statistically linear relationship b/w x & y



if $\beta_1 = 0$

- β_1 assumed to follow normal distribution

- Sampling distribution of β_1 follows t-distribution with n-1 dof

- Test if $\beta_1 = 0$ or not

$H_0 : \beta_1 = 0$ (no relationship b/w X & Y)
$H_a : \beta_1 \neq 0$ (there is a relationship b/w x & Y)

- Standard error of a statistic is the standard deviation of its sampling distribution

- SE of estimate: SD of sampling distribution of residuals

- If a sampled & calculated value of $\hat{\beta}_1 = 1$, how do we know if it is far enough from 0 to be statistically significant?

  - If $SE(\hat{\beta}_1) = 0.2$, then $\hat{\beta}_1 = 0.2 \times 5 =$ 5 standard errors away from 0

  - If $SE(\hat{\beta}_1) = 2$, then $\hat{\beta}_1 = 2 \times 0.5 =$ 0.5 standard errors away from 0

- How do we know no. of standard errors away from 0 that is enough to classify as statistically significant?

  - Critical value: cutoff no. of SEs needed for sample coefficient $\hat{\beta}_1$ to be statistically significant

- Equation for SE of estimates or SE of residuals

SS of residuals

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}}$$

2 dof lost from n samples due to $\hat{\beta}_0$ & $\hat{\beta}_1$

- Standard error of $\hat{\beta}_1$ (regression coefficient)

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X}_i)^2}}$$

Hypothesis Test

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

- t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_e(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{S_e(\hat{\beta}_1)}$$

Test for Overall Model : Analysis of Variance (F-test)

- Analysis of Variance (ANOVA) — test if statistically significant
- For SLR, H_0 and H_a of F-test same as t-test (same p-value)

H_0 : all regression coefficients are 0
H_a: not all regression coefficients are 0

- F statistic given by

$$\boxed{F \text{ statistic} = \frac{MSR}{MSE} = \frac{SSR}{\left(\frac{SSE}{n-2}\right)}}$$

- check p-value from table

Z-score

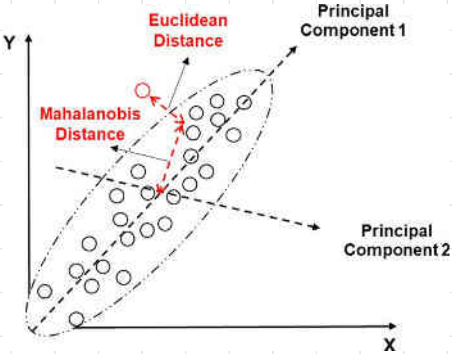
$$z = \frac{\hat{y}_i - \bar{y}}{\sigma_y}$$

Distance Measures

1. Mahalanobis Distance

- Distance between x_i and centroid of Y
- If distance value > χ^2 test critical value : outlier

https://youtu.be/3IdvoI8O9hU



- Euclidean distance: $D(\mu, x)$ — used commonly

mean

- $\sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$

- Very nice vid & channel: https://youtu.be/xc_X9GFVuVU (explains influence, leverage)

2. Cook's Distance

- How much $\hat{Y}$ changes when particular X_i excluded from sample for estimation of regression parameters

$$\boxed{D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1)\, MSE}}$$

TI, pg 250

where D_i is the Cook's distance measure for i^{th} observation, k is the number of predictors in the model, $\hat{Y}_j$ is the predicted value of j^{th} observation including i^{th} observation, $\hat{Y}_{j(i)}$ is the predicted value of j^{th} observation after excluding i^{th} observation from the sample, MSE is the Mean-Squared-Error. A Cook's distance value of more than 1 indicates highly influential observation.

- How much predicted value changes without a particular observation

3. Leverage Value

- Influence of an observation on overall fit of regression function

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

4. DFFit and DFBeta

- DFFit: change in $\hat{Y}_i$ when case i removed from the dataset
- DFBeta: change in regression coefficient values when obs i removed from dataset

Sum of Squared Errors

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})$$

- Regression model: minimise SSE

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{\bar{X}\sum_{i=1}^{n}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

GRADIENT DESCENT

- Hypothesis h for SLR

$$h_\theta = \theta_0 + \theta_1 x$$

- Find best fit line (alternative to OLS)
- Theory in ML

- Minimising the cost function by taking steps in the opposite direction of the gradient

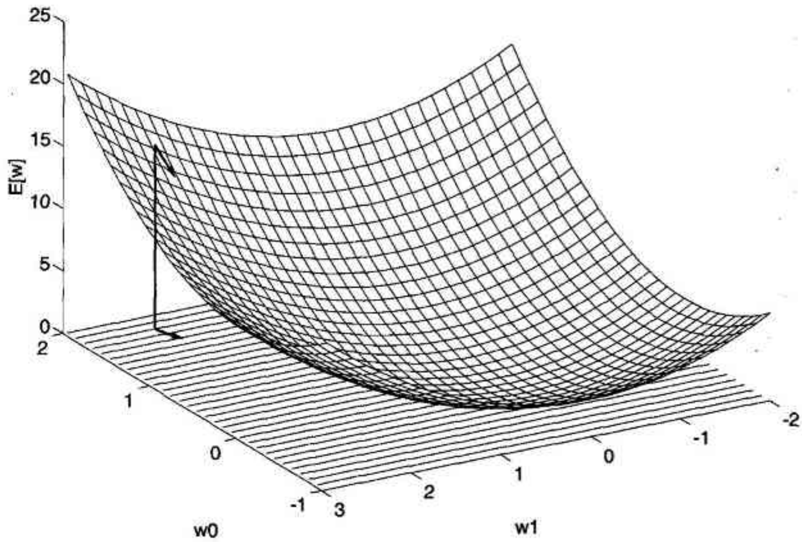


FIGURE 4.4
Error of different hypotheses. For a linear unit with two weights, the hypothesis space H is the w_0, w_1 plane. The vertical axis indicates the error of the corresponding weight vector hypothesis, relative to a fixed set of training examples. The arrow shows the negated gradient at one particular point, indicating the direction in the w_0, w_1 plane producing steepest descent along the error surface.

- Apply GD to $y = mx + c$
- Initially, let $m = 0$ and $c = 0$ (usually we never start with 0 but for SLR it is okay)
- Let LR = L (like 0.0001)
- Let E = loss function = $\frac{1}{n}$ SSE = $\frac{1}{n} \sum_{i=1}^{n} (\bar{y} - y_i)^2$
- Calculate $\frac{\partial E}{\partial m}$ and plug in x, y, m and c to obtain D_m

$$D_m = \frac{\partial E}{\partial m} = \frac{1}{n} \sum_{i=0}^{n} 2(\bar{y} - (mx_i + c))(-x_i)$$

$$D_m = \frac{-2}{n} \sum_{i=0}^{n} (\bar{y} - y_i) x_i$$

- Calculate $\frac{\partial E}{\partial c} = D_c$

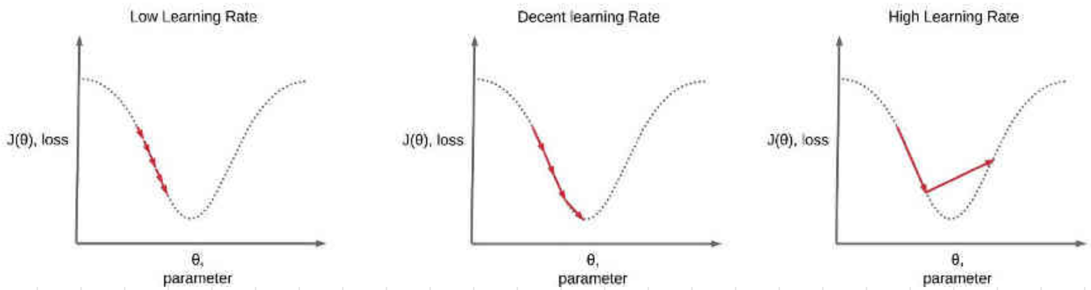
$$D_c = \frac{\partial E}{\partial c} = \frac{1}{n} \sum_{i=0}^{n} 2(\bar{y} - (mx_i + c))$$

$$D_c = \frac{2}{n} \sum_{i=0}^{n} (\bar{y} - y_i)$$

- Update values of m and c using D_m & D_c

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$



Source: deeplearning wizard

Multiple Linear Regression

- Functional form (can contain $\beta_i x_1^2$, $\beta_i x_1 x_2$ etc.)

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

- Matrix Representation of terms

$$\boldsymbol{Y} = \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_N \end{bmatrix}}_{N \times 1} \qquad \boldsymbol{X} = \underbrace{\begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & \cdots & X_{1,k} \\ 1 & X_{2,1} & X_{2,2} & & & X_{2,k} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ 1 & X_{N,1} & X_{N,2} & & & X_{N,k} \end{bmatrix}}_{N \times (k+1)}$$

$$\beta = \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}}_{(k+1) \times 1} \qquad \varepsilon = \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_N \end{bmatrix}}_{N \times 1}$$

- Matrix Representation of functional form

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & & x_{k1} \\ 1 & x_{12} & x_{22} & & x_{k2} \\ & & & & \\ & & & & \\ 1 & x_{1n} & x_{2n} & & x_{kn} \end{bmatrix} \bullet \begin{bmatrix} \beta_0 \\ \beta_1 \\ \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

OLS Estimation

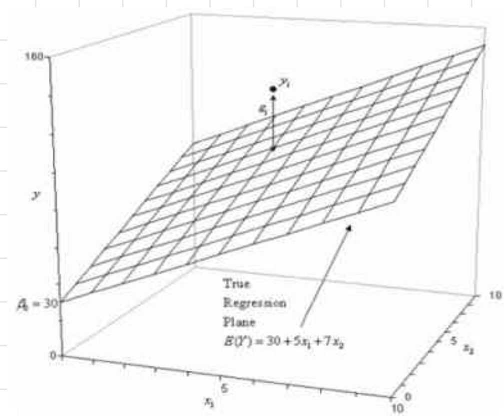
- Assumptions

1. The regression model is linear in parameter.
2. The explanatory variable, X_i, is assumed to be non-stochastic (that is, X_i is deterministic).
3. The conditional expected value of the residuals, $E(\varepsilon_i|X_i)$, is zero.
4. In a time series data, residuals are uncorrelated, that is, $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
5. The residuals, ε_i, follow a normal distribution.
6. The variance of the residuals, $Var(\varepsilon_i|X_i)$, is constant for all values of X_i. When the variance of the residuals is constant for different values of X_i, it is called **homoscedasticity**. A non-constant variance of residuals is called **heteroscedasticity**.
7. There is no high correlation between independent variables in the model (called **multi-collinearity**). Multi-collinearity can destabilize the model and can result in incorrect estimation of the regression parameters.

$$\hat{Y} = X\hat{\beta}$$

Multiply by X^T to left

$$X^T Y = X^T X \hat{\beta}$$



$X^T X$ is square & invertible

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

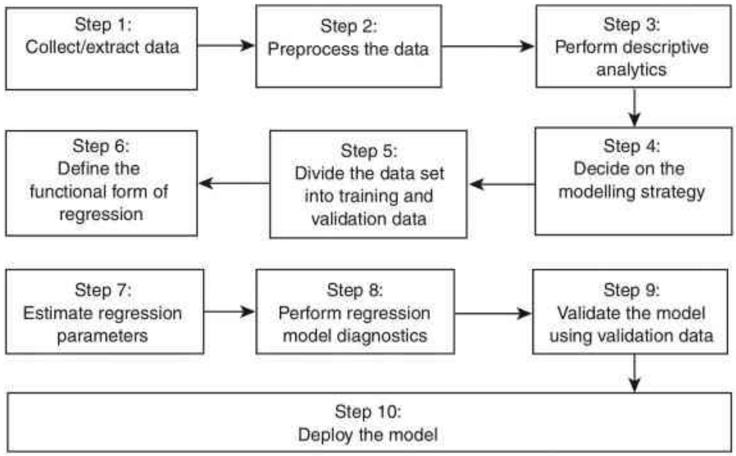
Estimated value of $\hat{Y}$

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{} Y$$

$$\hat{Y} = HY$$

H = hat matrix / influence matrix

Framework for Building MLR



(read steps from T1)

FIGURE 10.1 Framework for building multiple linear regression (MLR).

T1

Estimate $\hat{\beta}$

- OLS provides Best Linear Unbiased Estimate (BLUE)
- OLS fits polygon such that SSE is minimum

Model Diagnostics

- F-test – overall significance of model
  t-test – significance of each variable
- Presence of multi-colinearity : Variance Inflation Factor (VIF)
  (to drop attributes)
- Adjusted R^2 for MLR as R^2 normally increases with dimensionality

- Mean absolute percentage error

$$\sum_{i=1}^{k} \frac{1}{k} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\%$$

k = no. of validation cases

- RMS error

$$\sqrt{\sum_{i=1}^{k} \frac{1}{k} (Y_i - \hat{Y}_i)^2}$$

Part (Semi-Partial) Correlation and Regression Model Building

- Increase in R^2 when a new variable is added is given by the square of the semi-partial correlation of the newly added variable with dependent variable Y

- Model with two independent variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

1. Partial Correlation

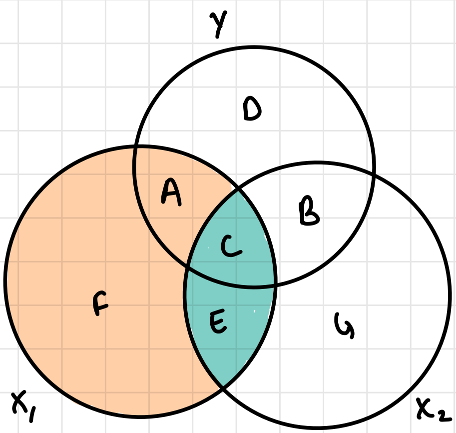
- Correlation between X_1 & Y when X_2 is constant

- Partial correlation of X_1 & Y when X_2 is constant

$$r_{YX_1,X_2} = \frac{r_{YX_1} - r_{YX_2} \times r_{X_1X_2}}{\sqrt{(1 - r^2_{YX_2}) \times (1 - r^2_{X_1X_2})}}$$

2. Semi-Partial Correlation

- Also called part correlation
- Relationship between Y & X_1 when influence of X_2 removed only from X_1 and not from Y
- Removing C and E from X_1 in the Venn diagram below



- When influence of X_2 removed from X_1

$$sr_{YX_1,X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{(1 - r^2_{X_1X_2})}}$$

Q: The cumulative television rating points (CTRP) of a television program, money spent on promotion (denoted as P), and the advertisement revenue (in Indian rupees denoted as R) generated over one-month period for 38 different television programs is provided in Table 10.1. Develop a multiple linear regression model to understand the relationship between the advertisement revenue (R) generated as response variable and promotions (P) and CTRP as predictors.

TABLE 10.1 Data on advertisement revenue (*R*) of programs along with *CTRP* and *P*

Serial	*CTRP*	*P*	*R*	Serial	*CTRP*	*P*	*R*
1	133	111600	1197576	20	156	104400	1326360
2	111	104400	1053648	21	119	136800	1162596
3	129	97200	1124172	22	125	115200	1195116
4	117	79200	987144	23	130	115200	1134768
5	130	126000	1283616	24	123	151200	1269024
6	154	108000	1295100	25	128	97200	1118688
7	149	147600	1407444	26	97	122400	904776
8	90	104400	922416	27	124	208800	1357644
9	118	169200	1272012	28	138	93600	1027308
10	131	75600	1064856	29	137	115200	1181976
11	141	133200	1269960	30	129	118800	1221636
12	119	133200	1064760	31	97	129600	1060452
13	115	176400	1207488	32	133	100800	1229028
14	102	180000	1186284	33	145	147600	1406196
15	129	133200	1231464	34	149	126000	1293936
16	144	147600	1296708	35	122	108000	1056384
17	153	122400	1320648	36	120	194400	1415316
18	96	158400	1102704	37	128	176400	1338060
19	104	165600	1184316	38	117	172800	1457400

MLR Model

$$R = \beta_0 + \beta_1 \, CTRP + \beta_2 P$$

TABLE 10.2 Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.912[a]	0.832	0.822	57548.382

[a]Predictors: (Constant), *P*, *CTRP*.

Adjusted R^2

$$Adjusted\ R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$

n = no. of datapoints
k = no. of independent vars
R^2 = R^2 value

more later (pg 54)

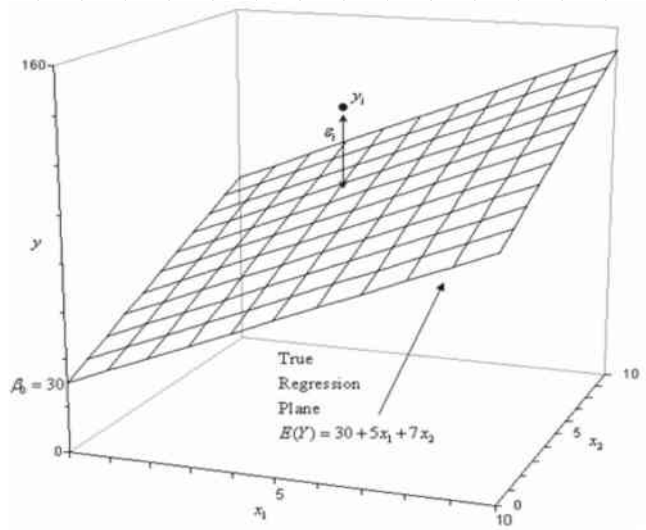
more on it later

TABLE 10.3 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Constant	41008.840	90958.920		0.451	0.655
	CTRP	5931.850	576.622	0.732	10.287	0.000
	P	3.136	0.303	0.736	10.344	0.000

$$R = 41008.840 + 5931.850 \, CTRP + 3.136 \, P$$

Visualisation of MLR



Partial Regression Coefficients

1. R and CTRP

$$R = \alpha_0 + \alpha_1 \times CTRP + \varepsilon_1$$

- Independent var decided from domain knowledge

TABLE 10.4 Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.564	0.318	0.299	114293.708

TABLE 10.5 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	625763.106	141522.635		4.422	0.000
	CTRP	4569.214	1114.912	0.564	4.098	[illegible]

2. P and CTRP

$$P = \delta_0 + \delta_1 \times CTRP + \varepsilon_2$$

TABLE 10.6 Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.228	0.052	0.026	31635.39950

TABLE 10.7 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	186456.659	39172.105		4.760	0.000
	CTRP	−434.495	308.597	−0.228	−1.408	0.168

3. ε_1 and ε_2

$$\varepsilon_1 = \eta_0 + \eta_1 \times \varepsilon_2 + \varepsilon_3$$

TABLE 10.8 Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.868	0.754	0.747	56743.46998

TABLE 10.9 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−1.614E-010	9205.006		0.000	1.000
	Unstandardized Residual	3.136	0.299	0.868	10.491	0.000

Summary

TABLE 10.10 Model development in multiple linear regression

Model	Estimated Parameters Values	Model Interpretation
$R = \beta_0 + \beta_1 \times CTRP + \beta_2 \times P + \varepsilon$	$R = 41008.84 + 5931.85\ CTRP + 3.136 \times P$	Variation in R explained by $CTRP$ and P
$R = \alpha_0 + \alpha_1 \times CTRP + \varepsilon_1$	$R = 625763.106 + 4569.214 \times CTRP$	Variation in R explained by $CTRP$
$P = \delta_0 + \delta_1 \times CTRP + \varepsilon_2$	$P = 186456.659 - 434.495 \times CTRP$	Variation in P explained by $CTRP$
$\varepsilon_1 = \eta_0 + \eta_1 \times \varepsilon_2$	$\varepsilon_1 = 3.136 \times \varepsilon_2$ (The value of η_1 is same as that of β_2)	ε_1 is variation in R not explained by $CTRP$; ε_2 is variation in P not explained by $CTRP$

- Every new variable added to the model is partialled out from other independent variables and regressed with the partialled out dependent variable

STANDARDISED REGRESSION COEFFICIENTS

- Regression model built on standardised dependent & independent variables

- Standardised Beta

$$\hat{\beta} \times \frac{S_{x_i}}{S_y}$$

- Interpretation: for one SD change in explanatory (x) variable, no. of SDs response (y) variable changes by

One SD change of CTRP, 0.732 SDs change of y

TABLE 10.11 Standardized regression coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	41008.840	90958.920		0.451	0.655
	CTRP	5931.850	576.622	0.732	10.287	0.000
	P	3.136	0.303	0.736	10.344	0.000

Regression Models with Qualitative Variables

- Preprocess categorical vars using dummy variables

Q: The data in Table 10.12 provides salary and educational qualifications of 30 randomly chosen people in Bangalore. Build a regression model to establish the relationship between salary earned and their educational qualifications.

TABLE 10.12 Education versus salary

S. No.	Education[a]	Salary	S. No.	Education	Salary	S. No.	Education	Salary
1	1	9800	11	2	17200	21	3	21000
2	1	10200	12	2	17600	22	3	19400
3	1	14 200	13	2	17650	23	3	18800
4	1	21000	14	2	19600	24	3	21000
5	1	16500	15	2	16700	25	4	6500
6	1	19210	16	2	16700	26	4	7200
7	1	9700	17	2	17500	27	4	7700
8	1	11000	18	2	15000	28	4	5600
9	1	7800	19	3	18500	29	4	8000
10	1	8800	20	3	19700	30	4	9300

[a] 1 — High school, 2 — Under-graduate, 3 — Post-graduate and 4 — None.

- Use 3 dummy variables and one-hot encoding

TABLE 10.13 Pre-processed data (sample)

Observation	Education	Pre-Processed data			Salary
		High School (*HS*)	Under-Graduate (*UG*)	Post-Graduate (*PG*)	
1	1	1	0	0	9800
11	2	0	1	0	17200
19	3	0	0	1	18500
27	4	0	0	0	7700

$$Y = \beta_0 + \beta_1 \times HS + \beta_2 \times UG + \beta_3 \times PG$$

- Base category: 000 for None

TABLE 10.14 Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t-value	p-value
1	(Constant)	7383.333	1184.793		6.232	0.000
	High-School (*HS*)	5437.667	1498.658	0.505	3.628	0.001
	Under-Graduate (*UG*)	9860.417	1567.334	0.858	6.291	0.000
	Post-Graduate (*PG*)	12350.000	1675.550	0.972	7.371	0.000

$$Y = 7383.333 + 5437.667\ HS + 9860.417\ UG + 12350.000\ PG$$

Interaction Variables

- Product of two variables (eg: $X_1 X_2$)
- Usually product of categorical and continuous

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Q: The data in Table 10.15 provides salary, gender, and work experience (WE) of 30 workers in a firm. In Table 10.15, gender = 1 denotes female and 0 denotes male and WE is the work experience in number of years. Build a regression model by including an interaction variable between gender and work experience. Discuss the insights based on the regression output.

TABLE 10.15 Data on salary, gender, and work experience (WE)

S. No.	Gender	WE	Salary	S. No.	Gender	WE	Salary
1	1	2	6800	16	0	2	22100
2	1	3	8700	17	0	1	20200
3	1	1	9700	18	0	1	17700
4	1	3	9500	19	0	6	34700
5	1	4	10100	20	0	7	38600
6	1	6	9800	21	0	7	39900
7	0	2	14500	22	0	7	38300
8	0	3	19100	23	0	3	26900
9	0	4	18600	24	0	4	31800
10	0	2	14200	25	1	5	8000
11	0	4	28000	26	1	5	8700
12	0	3	25700	27	1	3	6200
13	0	1	20350	28	1	3	4100
14	0	4	30400	29	1	2	5000
15	0	1	19400	30	1	1	4800

$$Y = \beta_0 + \beta_1 \times \text{Gender} + \beta_2 \times WE + \beta_3 \times \text{Gender} \times WE$$

TABLE 10.16 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13443.895	1539.893		8.730	0.000
	Gender	–7757.751	2717.884	–0.348	–2.854	0.008
	WE	3523.547	383.643	0.603	9.184	0.000
	Gender*WE	–2913.908	744.214	–0.487	–3.915	0.001

$$Y = 13443.895 \underbrace{- 7757.751 \times \text{Gender}} + 3523.547\, WE - 2913.908 \times WE \times \text{Gender}$$

Gender female = 1
male = 0

Gender ↑ (female), salary ↓ by 7757.751

Regression Model Diagnostics

- F-test – overall significance of model
  t-test – significance of each variable

- Presence of multi-colinearity : Variance Inflation Factor (VIF)
  (to drop attributes)

- Mean absolute percentage error

$$\sum_{i=1}^{k} \frac{1}{k} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

k = no. of validation cases

- RMS error

$$\sqrt{\sum_{i=1}^{k} \frac{1}{k} (y_i - \hat{y}_i)^2}$$

Adjusted R^2

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

$$\boxed{\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}}$$

k = no. of independent vars

t-Test — Statistical Significance of Individual Variables

- Estimate of $\hat{\beta}$ (pg 42)

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- Residuals follow normal distribution (assumption for MLR)
  $\Rightarrow$ Y follows ND
  $\Rightarrow$ $\hat{\beta}$ follows ND

- SD of β estimated from the sample $\Rightarrow$ t-test used

- Hypothesis test

  H_0: no relationship between X_i and Y
  H_a: there is a relationship between X_i and Y

- Or

  $H_0: \beta_i = 0$
  $H_a: \beta_i \neq 0$

- $$t = \frac{\hat{\beta}_i - 0}{S_e(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

F-test — Statistical Significance of Overall Model

- Check statistical significance of overall model with α
- Conduct residual analysis (pg 20)
- Check for presence of multi-colinearity (strong correlation b/w independent variables)
- Hypothesis test

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_a: \text{not all } \beta_i\text{'s are } 0$$

- F statistic

$$F = \frac{MSR}{MSE} \quad \text{(pg 32)}$$

Partial F-Test

- Assume for dataset of N observations, a full model (k ind. vars) and a reduced model (r ind. vars) are defined (r < k)
- Full model

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

- Reduced model

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_r X_r$$

- Objective of Partial F-test : check if additional variables $x_{r+1}, x_{r+2}, \dots, x_k$ are statistically significant

- Hypothesis test

$$H_0: \beta_{r+1} = \beta_{r+2} = \dots = \beta_k = 0$$

$$H_a: \text{not all } \beta_{r+1}, \dots, \beta_k \text{ are } 0$$

- Partial F statistic

$$\boxed{\text{Partial } F = \left(\frac{(SSE_R - SSE_F)/(k-r)}{MSE_F} \right)}$$

SSE_R : SSE in reduced model
SSE_F : SSE in full model
MSE_F : MSE in full model

$$\text{Partial } F = \frac{(R^2_{\text{full}} - R^2_{\text{reduced}}) \ / \ (k-r)}{(1 - R^2_{\text{full}}) \ / \ (N-k-1)}$$

Variance Inflation Factor

- Extent of multicolinearity

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Regression model b/w X_1 & X_2

$$X_1 = \alpha_0 + \alpha_1 X_2$$

- Variance inflation factor

$$VIF = \frac{1}{1 - R_{12}^2}$$

$(1 - R_{12}^2)$: tolerance

- Tolerance : $1 - R_{12}^2$
- In t-test, t-statistic $\frac{\hat{\beta}}{S_e(\hat{\beta})}$ may be deflated due to colinearity
- $S_e(\hat{\beta})$ inflated by $\sqrt{VIF}$, or t-statistic deflated by $\sqrt{VIF}$
- Actual t-value given as

$$\boxed{t_{actual} = \frac{\hat{\beta}}{S_e(\hat{\beta})} \times \sqrt{VIF}}$$

Handling Multi-Collinearity

- PCA — creates orthogonal components
- Advanced models: Ridge Regression, LASSO Regression

Auto-Correlation

- Time-series data

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

- Read TI 10.15 for explanation
- Under-estimation of p-value
- Presence of auto-correlation determined from Durbin-Watson Test

Durbin-Watson Test

- Let ρ be correlation between ε_t and ε_{t-1} (Pearson's)
- Hypothesis test

$$H_0: \rho = 0$$
$$H_a: \rho \neq 0$$

- DW statistic

need not memorize

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

- D lies between 0 & 4
- Critical values: D_L and D_U

1. If $D < D_L$, then the errors are positively correlated.
2. If $D > D_L$, then there is no evidence for positive auto-correlation.
3. If $D_L < D < D_U$, the Durbin–Watson test is inconclusive.
4. If $(4 - D) < D_L$, then errors are negatively correlated.
5. If $(4 - D) > D_U$, there is no evidence for negative auto-correlation.
6. If $D_L < (4 - D) < D_U$, the test is inconclusive.

Distance Measures

1. Mahalanobis Distance

- Distance between x_i and centroid of Y
- If distance value > χ^2 test critical value : outlier
- https://youtu.be/3IdvoI8O9hU (pg 36)

$$D_M(x_i) = \sqrt{(x_i - \mu_i)^T S^{-1} (x_i - \mu_i)}$$

- Takes spread into account (S^{-1} : covariance matrix)
- Helps find outliers

2. Cook's Distance

- Measures change in β when a sample is left out

$$D_i = \frac{(\hat{y}_j - \hat{y}_{j(i)})^T (\hat{y}_j - \hat{y}_{j(i)})}{(k+1) \times MSE}$$

3. Leverage Value

- Influence of an observation on overall fit

$$h_i = [H_{ii}] = X(X^T X)^{-1} X^T$$

$$h_i = \frac{\text{Mahalanobis Dist}^2}{N-1} + \frac{1}{N}$$

4. DFFit and DFBeta

- DFFit : diff in fitted value when observation is removed
- SDFFit : standardized DFFit

$$DFFit = \hat{y}_i - \hat{y}_{i(i)}$$

- $\hat{y}_i$: prediction of ith value, including ith observation
- $\hat{y}_{i(i)}$: prediction of ith value, excluding ith observation
- SDFFit

$$SDFFit = \frac{\hat{y}_i - \hat{y}_{i(i)}}{S_e(i)\sqrt{h_i}}$$

$S_e(i)$ is the standard error of estimate of the model after removing i^{th} observation and h_i is the i^{th} diagonal element in the hat matrix. The threshold for DFFIT is defined using **Standardized DFFIT** (SDFFIT). The absolute value of SDFFIT should be less than $2\sqrt{(k+1)/N}$.

- DFBeta

$$DFBeta_i(j) = \hat{\beta}_j - \hat{\beta}_{j(i)}$$

where $DFBETA_i(j)$ is the change in the regression coefficient for independent variable j when observation i is excluded. $\hat{\beta}_j$ is the estimated value of j^{th} regression coefficient including i^{th} observation, $\hat{\beta}_{j(i)}$ is the estimated value of j^{th} regression coefficient after excluding i^{th} observation from the sample.

- SDFBeta

$$SDFBeta_i(j) = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{Se(\hat{\beta}_{j(i)})}$$

Variable Selection in Regression Model Building

(a) Forward Selection

- k independent variables in dataset
- One variable added at every step

Step 1

- Start with 0 variables
- Calculate correlation b/w all X's and Y

Step 2

- Develop SLR by adding variable with highest ρ

$$Y = \beta_0 + \beta_1 X_i$$

- Create new model $Y = \alpha_0 + \alpha_1 X_i + \alpha_2 X_j \quad (j \neq i)$
  (there are $k-1$ models)

- Conduct partial F-test to check if X_j is statistically significant at α

Step 3

- Add X_j from step 2 with smallest p-value based on partial F-test (if p-value $< \alpha$)

Step 4

- Repeat step 3 until smallest p-value is $> \alpha$ or all variables are exhausted

(b) Backward Selection

- Remove one variable at a time using F-test

Step 1

- Assume model is MLR with n independent variables (full model)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Step 2

- Remove one variable X_i from the model (there will be k such models)
- Perform partial F-test between the models in step 1 and step 2

Step 3

- Remove the variable with largest p-value if p-value $> \alpha$

Step 4

- Repeat until largest p-value $< \alpha$ or no variables in model

(c) Stepwise Regression

- Combination of forward and backward
- Entering criteria (α) based on smallest p-value of partial F-test
- Exiting criteria (β) based on largest p-value ($\alpha < \beta$)

TABLE 10.21 Variables entered/removed

Model	Variables Entered	Variables Removed	Method
1	*P*	.	Stepwise (Criteria: Probability-of-*F*-to-enter <= 0.050, Probability-of-*F*-to-remove >= 0.100).
2	*CTRP*	.	Stepwise (Criteria: Probability-of-*F*-to-enter <= 0.050, Probability-of-*F*-to-remove >= 0.100).

Avoiding Overfitting – Mallow's C_p

$$C_p = \left(\frac{SSE_p}{MSE_{full}}\right) - (N - 2p)$$

where SSE_p is the sum of squared errors with p parameters in the model (including constant), MSE_{full} is the mean squared error with all variables in the model, N is the number of observations, p is the number of parameters in the regression model including constant.

Transformations

- Deriving new dependent/independent variables to identify correct functional form of LR

- Transformation helps with

  1. Poor fit (low R^2)
  2. Pattern in residual analysis (pg 20)
  3. Residuals do not follow normal dist
  4. Residuals are not homoscedastic

Q: Table 10.28 shows the data on revenue generated (in million of rupees) from a product and the promotion expenses (in million of rupees). Develop an appropriate regression model.

TABLE 10.28 Data on revenue generated and promotion expenses

S. No.	Revenue in Millions	Promotion Expenses	S. No.	Revenue in Millions	Promotion Expenses
1	5	1	13	16	7
2	6	1.8	14	17	8.1
3	6.5	1.6	15	18	8
4	7	1.7	16	18	10
5	7.5	2	17	18.5	8
6	8	2	18	21	12.7
7	10	2.3	19	20	12
8	10.8	2.8	20	22	15
9	12	3.5	21	23	14.4
10	13	3.3	22	7.1	1
11	15.5	4.8	23	10.5	2.1
12	15	5	24	15.8	4.75

Y = Revenue in Millions
X = Promotion Expenses

Scatterplot

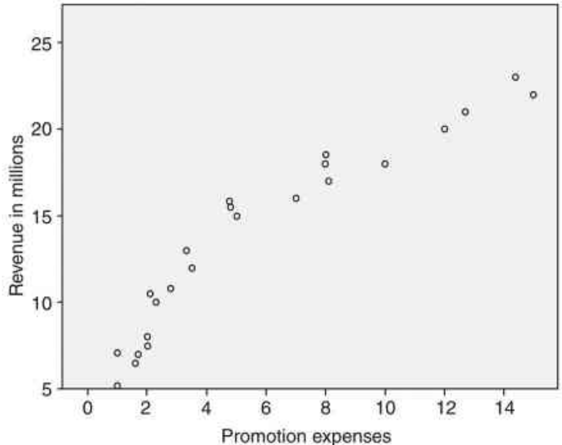


FIGURE 10.9 Scatter plot between promotion expenses and revenue in millions.

- Does not look linear. Consider

$$Y = \beta_0 + \beta_1 X$$

- Model summary

TABLE 10.29 Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.940	0.883	0.878	1.946

TABLE 10.30 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.831	0.650		10.516	0.000
	Promotion Expenses	1.181	0.091	0.940	12.911	0.000

- Residual plot

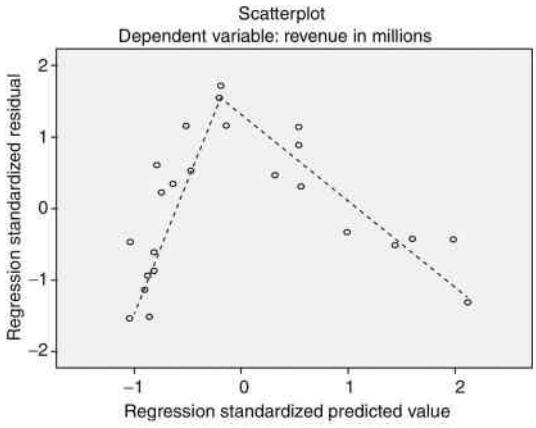


FIGURE 10.10 Residual plot.

- Transform into

$$Y = \beta_0 + \beta_1 \ln(X)$$

TABLE 10.31 Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.980	0.960	0.959	1.134

TABLE 10.32 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.439	0.454		9.771	0.000
	ln (X)	6.436	0.279	0.980	23.095	0.000

- New residuals - no pattern

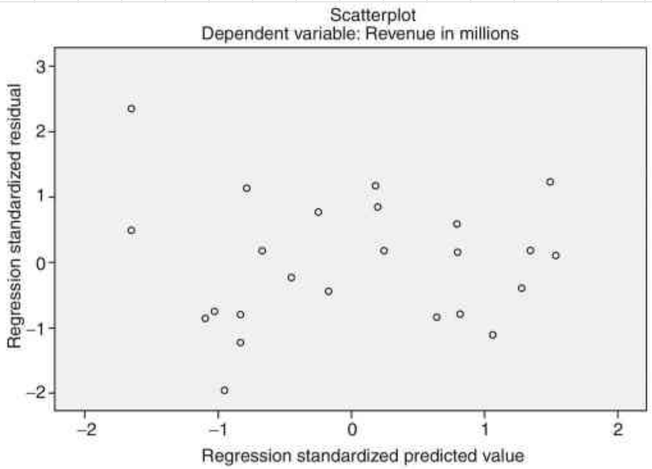


FIGURE 10.11 Residual plot for the model $Y = \beta_0 + \beta_1 \ln(X)$.

- How to decide on transformation?

TUKEY & MOSTELLER BULGING RULE

- To find correct functional form of regression

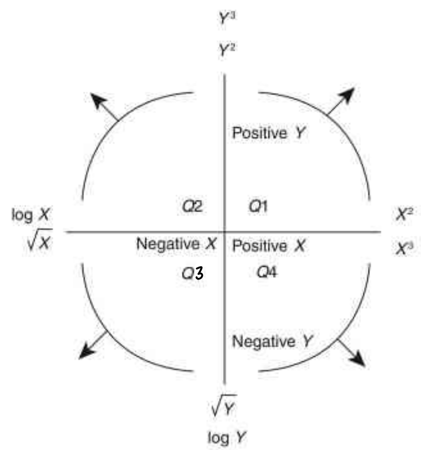


FIGURE 10.12 Tukey's Bulging Rule (adopted from Tukey and Mosteller, 1977).

- Suggested transformations based on the shape of the scatterplots as identified using the quadrants

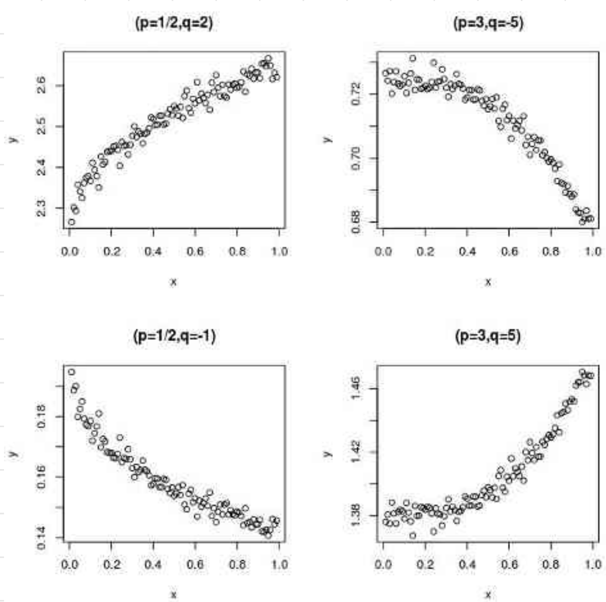
TABLE 10.33 Tukey's rule for transformations

Shape of Scatter Plot	Suggested Transformation for X	Suggested Transformation for Y
Q1 (X and Y positive)	X^p where $p > 1$ (e.g. X^2, X^3, etc.)	Y^q where $q > 1$(e.g. Y^2, Y^3, etc.)
Q2 (X negative and Y positive)	X^p where $p < 1$ (e.g., In(X), $\sqrt{X}$, etc.)	Y^q where $q > 1$ (e.g. Y^2 and Y^3 etc)
Q3 (Both X and Y negative)	X^p where $p < 1$ (e.g. In(X), $\sqrt{X}$, etc.)	Y^q where $q < 1$ (e.g. In(Y), $\sqrt{Y}$, etc.)
Q4 (X positive and Y negative)	X^p where $p > 1$ (e.g. X^2, X^3, etc.)	Y^q where $q < 1$ (e.g. In(Y), $\sqrt{Y}$, etc.)

should remember

- Fig. 10.9, Scatterplot from Q2

- https://freakonometrics.hypotheses.org/14967



Q&A (T1, pg 263)

1. Regression models cannot be used for
   (a) Analysing time-series data
   (b) Understanding association relationship
   (c) Understanding cause and effect relationship
   (d) All of the above

Explanation: (a) DW, autocorrelation,
(b) Functional form requires finding association relationship
(c) correlation ⇏ causation

2. The best simple linear regression model is the one for which
   (a) The R-square (coefficient) is the highest.
   (b) The residuals follow normal distribution.
   (c) The p-value corresponding to t-test is less than the significance value α.
   (d) The p-value corresponding to t-test is less than the significance value α and the residuals follow normal distribution and the residual are homoscedastic.

3. Which of the following equations are linear regression models?

(a) $Y = \beta_0 + \beta_1 X^2$ (b) $Y = \beta_0 + [1/(1+\beta_1)]\, X$ (c) $Y = \beta_0 + \beta_1 X$ (d) $\ln(Y) = \beta_0 + \beta_1 \ln(X)$

not

(wrt β_0, β_1

4. A high street jewellery shop uses a regression model $Y = -10.5 + 95 \times$ carat to predict the price of a diamond as a function of carat, where carat is the weight of the diamond. The value of β_0 is negative because:

(a) Regression model is incorrect since the value of diamond cannot take negative value.
(b) The regression models cannot be extrapolated beyond the range of the data used for building the model.
(c) The regression model is valid only for carat values greater than 0.1106 since the value of Y will be positive when carat is greater than 0.1106.
(d) The value of β_0 $(= -10.5)$ should be ignored while calculating the price of the diamond.

5. If the residuals do not follow normal distribution:

(a) The regression coefficient estimates are incorrect.
(b) The R-square values are incorrect.
(c) The standard error of estimate is incorrect.
(d) The t-test for the coefficient of the explanatory variable (β_1) is not valid.

6. If the correlation between a predictor variable and the outcome variable is 0.8, the proportion of variation in the outcome variable explained by the predictor variable is

(a) 0.9 (b) 0.72 (c) 0.89 (d) 0.64

$r = 0.8$ $R^2 = \frac{SSR}{SST}$ (pg 31)

7. Heteroscedasticity of the residual implies

(a) The variance of error for different values of the explanatory variables is different.
(b) The variance of error for different values of the explanatory variables is same.
(c) The variance of error decreases and the value of explanatory variable increases.
(d) The variance of error increases as the value of outcome variable increases.

8. In a model $\ln(Y) = \beta_0 + \beta_1 X$, the value of β_1 is

(a) Change in value of Y for unit change in value of X.
(b) Change in value of X for unit change in value of Y.
(c) Percentage change in value of X for unit change in value of Y.
(d) Percentage change in value of Y for unit change in value of X.

Note: 3^{rd} or 4^{th} Quadrant

9. Mahalanobis distance is a

(a) Measure of performance of the regression model.
(b) Measure of outlier.
(c) Measure of error.
(d) Measure of explained variation.

10. Transformation of outcome variable and predictor variable is used for

(a) Improving coefficient of determination.
(b) Removing heteroscedasticity
(c) Removing patterns in residual plot
(d) All of the above

Q: Professor Bell at Bellandur University, Bangalore believes that the cumulative grade point average (CGPA) of the students is negatively correlated with usage (measured in average minutes per day) of smart phones. Table 1 shows the CGPA and smart phone usage in minutes per day of 40 students.

(a) Calculate the Pearson correlation coefficient between CGPA and mobile phone usage of students.

(b) Conduct a hypothesis test at a = 0.01 to check whether CGPA and mobile phone usage are negatively correlated.

(c) Professor Bell believes that the correlation is less than −0.4. Conduct a hypothesis test at a = 0.1 to check whether the claim is correct.

Table.1: Data of CGPA and mobile phone usage (Average minutes per day)

CGPA	2.65	2.25	1.86	1.47	2.10	1.94	2.71	1.83	2.65	2.04
Phone Usage	75	89	65	136	95	103	74	109	7	98
CGPA	2.54	2.16	2.28	2.47	2.18	2.57	1.97	2.87	2.10	3.28
Phone Usage	60	93	88	81	92	78	102	70	95	89
CGPA	2.78	2.44	1.87	2.50	2.24	2.01	2.17	2.20	2.05	1.63
Phone Usage	72	82	107	80	89	100	92	91	98	123
CGPA	2.28	2.63	2.86	2.24	2.44	2.69	2.22	3.07	1.77	3.03
Phone Usage	88	76	70	89	82	74	90	65	113	66

(a) $$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_x S_y} = \frac{\Sigma xy - n\bar{x}\bar{y}}{S_x S_y}$$

$$= \frac{7786.37 - 10 \times 86.15 \times 2.326}{20.592 \times 0.407} = -0.8026$$

(b) $\alpha = 0.01$

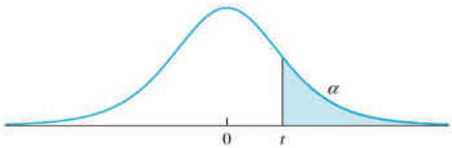
$H_0 : \rho \geq 0$
$H_a : \rho < 0$

dof = 38

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.8026}{0.0967} = -8.29$$

(Check for right tail 8.29)

TABLE A.3 Upper percentage points for the Student's *t* distribution



	α								
ν	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

critical point: between 2.423 and 2.438

$\therefore$ reject H_0 and accept H_a

(c) $H_0: \rho \geq -0.4$ $\alpha = 0.1$
$H_a: \rho < -0.4$

$$t = \frac{-0.8026 + 0.4}{0.0967} = -4.163$$

critical point (for right tail) : b/w 1.303 & 1.306

$\therefore$ we reject H_0

Q: Calculate the Spearman rank correlation between the corruption rank and literacy rank.

TABLE 3 Rank based on corruption (1 implies high corruption)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	1	2	3	4	5	6	7	8
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	9	10	11	12	13	14	15	16

TABLE 4. Rank based on literacy rate (1 implies high literacy)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	16	12	10	11	7	15	4	9
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	2	5	13	1	8	14	6	3

Conduct a hypothesis test to check whether corruption and literacy rate are negatively correlated at a = 0.05.

$$r = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

State	Rank corr	Rank lit	$\lvert D \rvert$ (diff)
Bihar	1	16	15
J&K	2	12	10

$r = -0.4588$ (negative correlation)

$H_0: \rho = 0$ $\alpha = 0.05$ $\alpha/2 = 0.025$

$H_a: \rho \neq 0$ dof = 14

$t = -1.932$

p-value for 2-tailed test is 2.145

$\therefore$ cannot reject H_0

Multivariate Regression (MvLR)

- Predict multiple dependent variables from multiple independent variables

$$\boxed{y_{ik} = b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij} + e_{ik}}$$

$i \in \{1, 2, \ldots, n\}$

$k \in \{1, 2, \ldots, m\}$

- $y_{ik} \in \mathbb{R}$ is the k-th real-valued response for the i-th observation
- $b_{0k} \in \mathbb{R}$ is the regression intercept for k-th response
- $b_{jk} \in \mathbb{R}$ is the j-th predictor's regression slope for k-th response
- $x_{ij} \in \mathbb{R}$ is the j-th predictor for the i-th observation
- $(e_{i1}, \ldots, e_{im}) \stackrel{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \mathbf{\Sigma})$ is a multivariate Gaussian error vector

Assumptions of MvLR

1. Relationship b/w x_j & y_k is linear
2. x_{ij} and y_{ik} are observed random variables (known constants)
3. $(e_{i1}, \ldots, e_{im}) \stackrel{\text{iid}}{\sim} N(0_m, \Sigma)$ is an unobserved random vector
4. $b_k = (b_{0k}, b_{1k}, \ldots, b_{pk})'$ for $k \in \{1, \ldots, m\}$ are unknown constants
5. $(y_{ik} | x_{i1}, \ldots, x_{ip}) \sim N(b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij}, \sigma_{kk})$ for each $k \in \{1, \ldots, m\}$

MvLR Model-Matrix Form

http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf

$$Y = XB + E$$

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ y_{31} & \cdots & y_{3m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_{01} & \cdots & b_{0m} \\ b_{11} & \cdots & b_{1m} \\ b_{21} & \cdots & b_{2m} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ e_{21} & \cdots & e_{2m} \\ e_{31} & \cdots & e_{3m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

Hat Matrix

$$\begin{aligned} \hat{Y} &= X\hat{B} \\ &= X(X^T X)^{-1} X^T Y \\ &= HY \end{aligned}$$

symmetric, idempotent

Bias-Variance Trade-off

- http://scott.fortmann-roe.com/docs/BiasVariance.html ← really good explanation
- n observations
- SLR with X & Y
- Normally distributed error term with variance σ^2

$$Y = X\beta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

- Error ε is ideally close to 0
- True value of β unknown
- Estimated as $\hat{\beta}$ such that SSR is minimum (OLS)

$$\text{Recall} \quad R^2 = \frac{SSR}{SST}$$

- Loss function (sum of squared errors)

$$L_{OLS}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i \hat{\beta})^2 = \|y - X\hat{\beta}\|^2$$

$$\hat{\beta}_{OLS} = (X^T X)^{-1} (X^T Y)$$

Bias

- Difference between true population parameter and expected sample estimator

$$\text{Bias}(\hat{\beta}_{OLS}) = E(\hat{\beta}_{OLS}) - \beta$$

- Measures accuracy

Variance

- Measures spread or uncertainty

$$Var(\hat{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1}$$

- Unknown error variance σ^2 is estimated from residuals as

$$\hat{\sigma}^2 = \frac{\varepsilon^T \varepsilon}{n-m} \qquad \varepsilon = y - X\hat{\beta}$$

$\varepsilon^T \varepsilon$: SSE of ε
n : no. of samples
m: no. of parameters being estimated $(\hat{\beta})$

Visualizing — Bull's Eye Model

source: http://scott.fortmann-roe.com/docs/BiasVariance.html

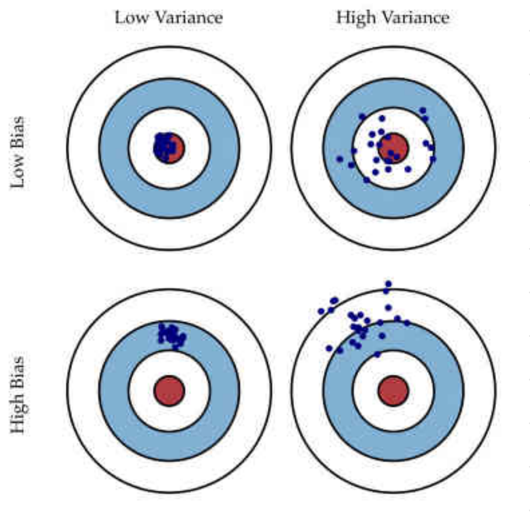


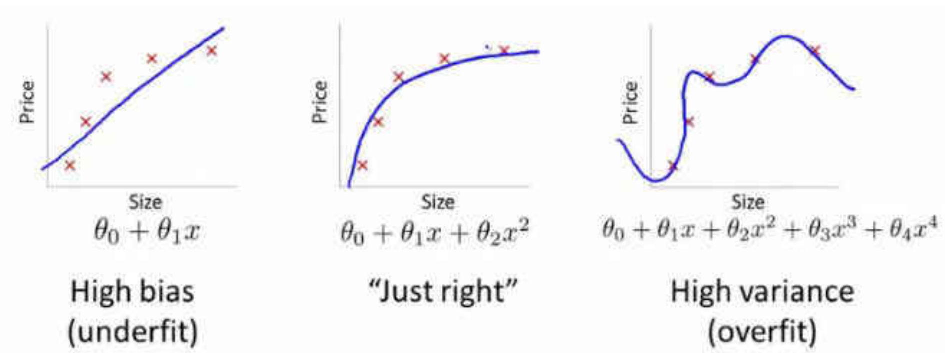
Fig. 1 Graphical illustration of bias and variance.

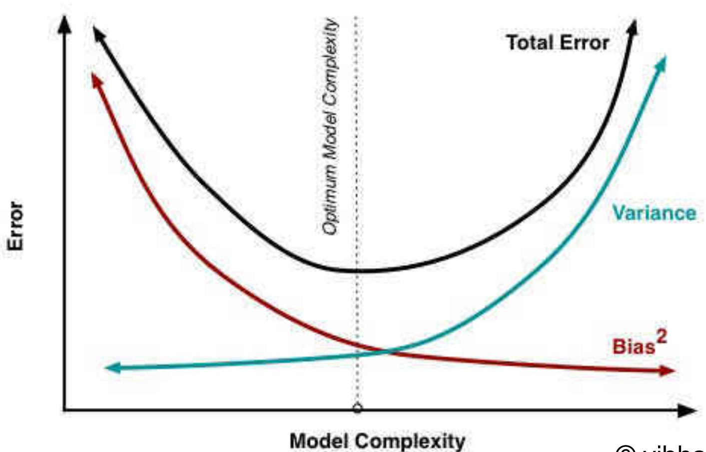
Model's Error in Terms of Bias and Variance

- Error decomposed into 3 parts
  (1) Error due to large variance
  (2) Error due to significant bias
  (3) Unexplained error

$$E(\varepsilon) = (E(X\hat{\beta}) - X\beta)^2 + E\left((X\hat{\beta}) - E(X\hat{\beta}))^2\right) + \sigma^2$$

$$= \text{bias}^2 + \text{variance} + \sigma^2$$





LASSO & RIDGE REGRESSION

- Reduce model complexity (read slides for details)

1. LASSO

- Least Absolute Shrinkage Selector Operator
- Same assumptions as MLR except normality of error
- Tends to zero out (remove) some features (feature selection)
- Uses L_1 norm or absolute values of coefficients scaled by shrinkage

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- λ is tunable

2. Ridge

- Shrinkage term added to objective SSE function
- $\lambda = 0$ has no effect, ← OLS $\lambda \to \infty$ regression coefficient estimates $\to 0$

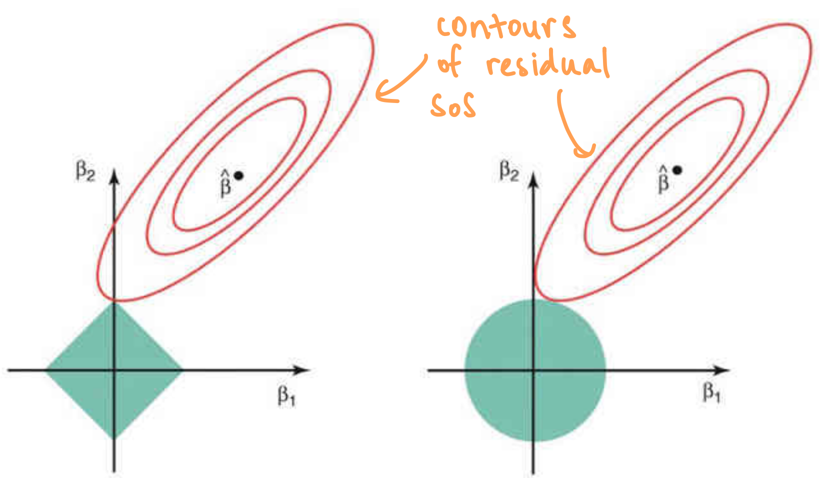
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} (\beta_j^2)$$

- Used when data suffers from multi-colinearity
- Must scale input

$$\bar{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}}$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \|\beta\|_1 \qquad \hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \|\beta\|_2^2$$

$$|\beta 1| + |\beta 2| \le S \qquad \beta 1^2 + \beta 2^2 \le S$$



Source: An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Source: Towards Data Science

L1 regularization L2 regularization

- Correlated vars have similar weights in Ridge
- One high & others close to 0 for Lasso (correlated vars)
- Achieve reduction in variance without increase in bias

Choice of Regularization Parameter

- How much bias acceptable to decrease variance?
- Choose λ so that AIC or BIC is smallest

- Estimate with many different values for λ and choose one that minimizes AIC or BIC

- Akaike Information Criterion (like R^2)

$$AIC_{ridge} = n \ln(\varepsilon^T \varepsilon) + 2 df_{rigde}$$ } want to min

- Bayesian Information Criterion (like adjusted R^2)

$$BIC_{ridge} = n \ln(\varepsilon^T \varepsilon) + 2\, df_{ridge} \ln(n)$$ } want to min

parameters

- df_{ridge} = no of degrees of freedom

3. Elastic Net Regression

- Combining L1 & L2 norms

$$\hat{\beta} = \underset{\beta}{\arg\min} \left(\|y - X\beta\|^2 + \underbrace{\lambda_2 \|\beta\|^2}_{\text{ridge}} + \underbrace{\lambda_1 \|\beta_1\|}_{\text{lasso}} \right)$$

POLYNOMIAL REGRESSION MODEL

http://users.stat.umn.edu/~helwig/notes/polyint-Notes.pdf

$$f(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$$

$$= \sum_{j=0}^{n} a_j x^j$$

Model Form

$$y_i = \sum_{j=0}^{p} b_j x_i^j + \varepsilon_i$$

b_0 = intercept and $x_0 = 1$

$\varepsilon_i \sim N(0, \sigma^2)$

Matrix Form

$$Y = XB + e$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ 1 & x_3 & x_3^2 & \cdots & x_3^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix}$$

PR and MLR

- x & x^n are not independent in PR
- $n > p$ to fit polynomial regression model
- Multicolinearity in PR (high power, high multicolinearity)
- One solution: mean-centering ($x^n \rightarrow x^n$ - mean(x^n))

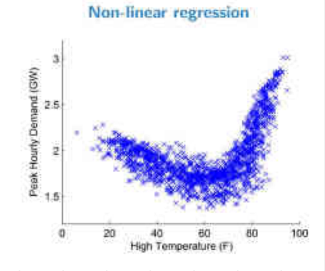
Solution 2: Orthogonal Polynomials

$$\left.\begin{aligned} z_0 &= a_0 \\ z_1 &= a_1 + b_1 x \\ z_2 &= a_2 + b_2 x + c_2 x^2 \\ z_3 &= a_3 + b_3 x + c_3 x^2 + d_3 x^3 \end{aligned}\right\} \text{orthogonal polynomials}$$

Such that $z_j^T z_k = 0$ for all $j \neq k$

Non-Linear Regression

- Detecting : theory, scatterplot, seasonality in data, insignificant β
- Can do incremental F-tests
- Two ways to construct non-linear features
  1. Explicitly – feature vector
  2. Implicitly – using kernels (read ML unit 2 - SVMs) – done by the model



Popular Models

1. Exponential model —— $y = ae^{bx}$
2. Power model —— $y = ax^b$
3. Saturation growth model —— $y = \dfrac{ax}{b+x}$

Logistic Regression

- For classification problems
- Odds: odds of an event with probability p occurring $= \frac{p}{1-p}$
  (ratio of success : failure)
- Eg: with binary gender, will a customer purchase a product or not

Gender	Purchase		Total
	Yes	No	
Female	159	106	265
Male	121	125	246

P(female purchase | customer is female) $= \frac{159}{265}$

P(not purchasing | customer is female) $= \frac{106}{265}$

odds(female purchasing) $= \frac{159}{106} \approx 1.5$

higher odds → higher chance of success

Odds Ratio

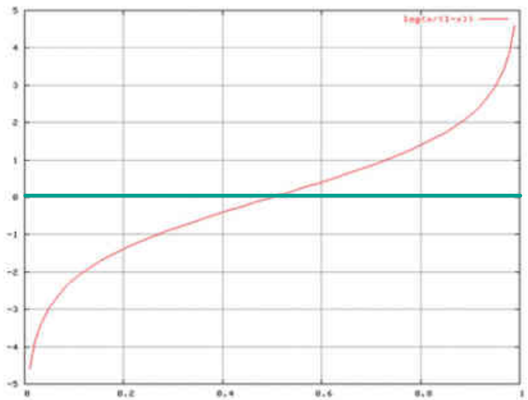
- Which group (male/female) has higher odds of success
- Odds ratio (female) $= \frac{\text{odds of successful female purchase}}{\text{odds of successful male purchase}}$

- Odds = $\frac{\pi}{1-\pi}$

- odds ratio = $\frac{(\pi_1)/(1-\pi_1)}{(\pi_0)/(1-\pi_0)}$

Logit Function

https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-f384152a33d1



$$Logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$$

$\pi \in [0, 1]$

When $\pi = 0.5$, $\ln\left(\frac{0.5}{0.5}\right) = 0$

- $y = \beta_0 + \beta_1 x$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

average (Logit(π))

Logistic Transformation

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$$

$$\frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 x}$$

$$\pi = (1-\pi)\, e^{\beta_0 + \beta_1 x}$$

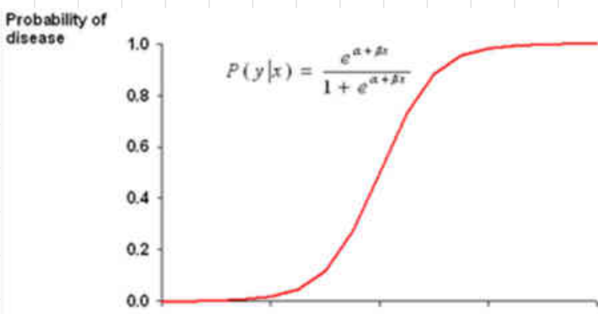
$$= e^{\beta_0 + \beta_1 x} - \pi e^{\beta_0 + \beta_1 x}$$

$$\pi\left(1 + e^{\beta_0 + \beta_1 x}\right) = e^{\beta_0 + \beta_1 x}$$

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 - \beta_1 x)}}$$

Logistic regression model

$$P(Y=1 \mid X = x) = \pi(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} = \text{sigmoid function}$$



$\beta = 0$ implies that $P(Y|x)$ is same for each value of x
$\beta > 0$ implies that $P(Y|x)$ is increases as the value of x increases
$\beta < 0$ implies that $P(Y|x)$ is decreases as the value of x increases

Likelihood Function for Binary Logistic Function

$$P(Y=1 \mid Z = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m) = \pi(z) = \frac{e^z}{1+e^z} \quad \text{pdf}$$

Probability Likelihood Function

$$P(Y_i) = \pi(z)^{y_i} (1-\pi(z))^{1-y_i}$$

Estimation of Parameters

- Assume n observations $Y_1, Y_2, \dots, Y_n$
- Likelihood function: joint probability $L = P(Y_1, Y_2, \dots, Y_n)$ for a specific $z_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_n x_{ni}$ is

$$L = P(Y_1, Y_2, \dots, Y_n) = \prod_{i=1}^{n} \pi(z_i)^{y_i} (1-\pi(z_i))^{1-y_i}$$

- Log-likelihood function

$$\ln(L) = \sum_{i=1}^{n} y_i \ln[\pi(z_i)] + \sum_{i=1}^{n} (1-y_i) \ln[1-\pi(z_i)]$$

- For simplicity, let $z_i = \beta_0 + \beta_1 x_i$

$$y_i \ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) + (1-y_i)\left(\ln\left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\right)$$

$$= y_i \ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) + (1-y_i)\left(\ln\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)\right)$$

$$= y_i(\beta_0+\beta_1 x_i) - \ln(1+e^{\beta_0+\beta_1 x_i})$$

$$LL = \ln(L) = \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) - \sum_{i=1}^{n} \ln(1+e^{\beta_0+\beta_1 x_i})$$

- Partial derivatives wrt β_0 & β_1

$$\frac{\partial LL}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0$$

$$\frac{\partial LL}{\partial \beta_1} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0$$

- No closed form solution $\longrightarrow$ gradient descent

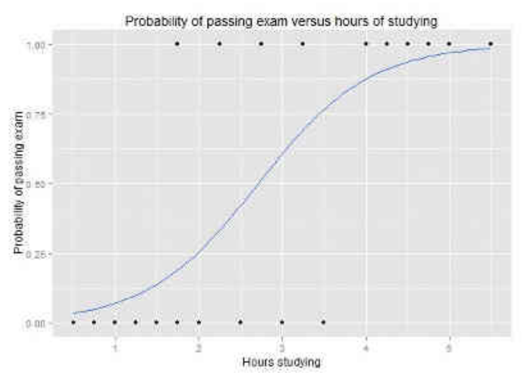
Interpretation of LR Coefficients

- β_1: change in ln(odds ratio) for unit change in X_1
- β_1 : change in odds ratio by e^{β_1}

$$\beta_1 = \ln\left(\frac{\pi(x+1) / (1-\pi(x+1))}{\pi(x) / (1-\pi(x+1))}\right)$$

Q: P(passing exam)

https://en.wikipedia.org/wiki/Logistic_regression



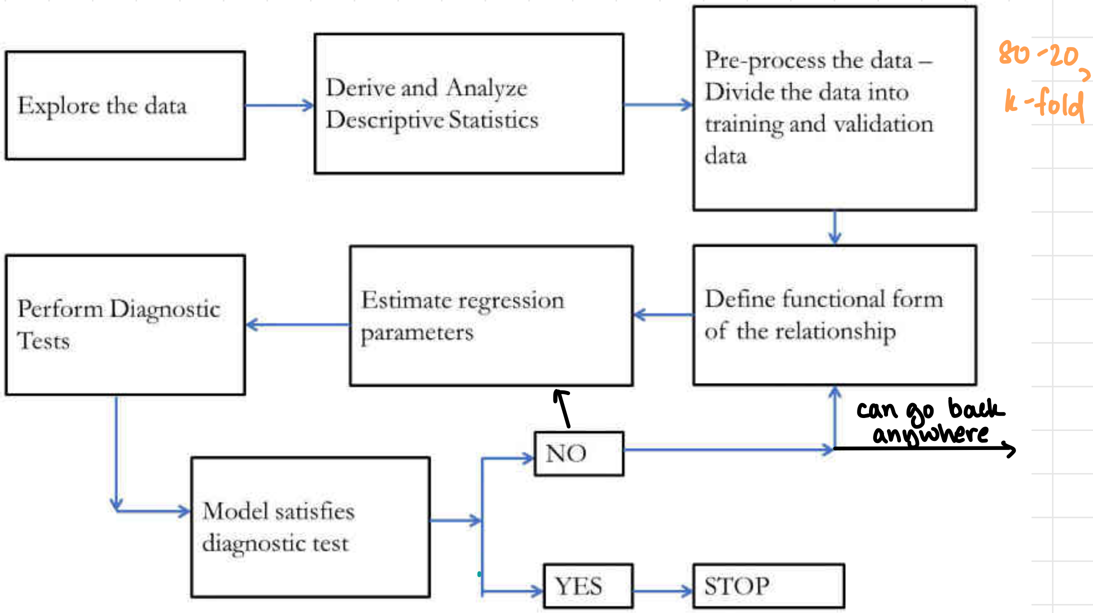
Hours of study	Passing exam		
	Log-odds	Odds	Probability
1	−2.57	0.076 ≈ 1:13.1	0.07
2	−1.07	0.34 ≈ 1:2.91	0.26
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97

	Coefficient	Std.Error	P-value (Wald)
Intercept	−4.0777	1.7610	0.0206
Hours	1.5046	0.6287	0.0167

$$z_i = -4.0777 + 1.5046\, x_i$$

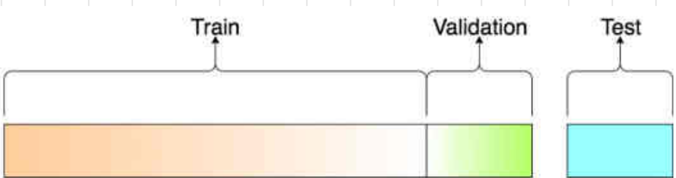
$$Y_i = \frac{1}{1+e^{-z_i}}$$

Logistic Regression Model Development



SPLITTING DATA

- **Training dataset:** sample used to fit model
- **Validation dataset:** sample used to provide unbiased evaluation of a model — tuning hyperparameters — becomes biased
- **Test dataset:** provides unbiased evaluation of final model — no further tuning



A visualization of the splits

- Dataset split ratio: depends on model ; can be 80-20, 70-30

CONFUSION MATRIX

		Predicted	
		Real email	Spam email
Actual	real email	TN	FP
	spam email	FN	TP

Accuracy

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall

- How many +ve cases caught (sencitivity) ; not missed
- True positive rate

$$\text{recall} = \frac{TP}{TP + FN}$$

Specificity

- How many -ve cases caught ; not missed

$$\text{specificity} = \frac{TN}{TN + FP}$$

Precision

- Correct positive cases out of predicted positive cases

$$\text{precision} = \frac{TP}{TP + FP}$$

F1 Score

- Harmonic mean of precision and recall

$$\text{F1 score} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

- Higher score → better (0 is worst, 1 is best)
- Only if precision and recall are 100%, F1 = 1

Q: Which model is better wrt class A?

(1)

		Prediction A	Prediction A'
Actual	A	TP 60	FN 34
	A'	FP 1	TN 12

(2)

		Prediction A	Prediction A'
Actual	A	TP 90	FN 4
	A'	FP 8	TN 5

$$\text{Accuracy}_1 = \frac{TP + TN}{TP + FP + TN + FN}$$

$$= \frac{72}{107} = 0.67$$

$$\text{Accuracy}_2 = \frac{TP + TN}{TP + FP + TN + FN}$$

$$= \frac{95}{107} = 0.89$$

$$\text{Precision}_1 = \frac{TP}{TP + FP}$$

$$= \frac{60}{61} = 0.98$$

$$\text{Precision}_2 = \frac{TP}{TP + FP}$$

$$= \frac{90}{98} = 0.92$$

$$\text{Recall}_1 = \frac{TP}{TP + FN}$$

$$= \frac{60}{94} = 0.64$$

$$\text{Recall}_2 = \frac{TP}{TP + FN}$$

$$= \frac{90}{94} = 0.96$$

$$\text{F1 Score}_1 = \frac{2 \times \frac{60}{94} \times \frac{60}{61}}{\frac{60}{94} + \frac{60}{61}}$$

$$= \frac{24}{31} = 0.77$$

$$\text{F1 Score}_2 = \frac{2 \times \frac{90}{98} \times \frac{90}{94}}{\frac{90}{98} + \frac{90}{94}}$$

$$= \frac{15}{16} = 0.94$$

$\therefore$ Model 2 is better

Multi-Class Confusion Matrix

		Predicted	
	A	B	C
Actual A	1	2	3
B	4	5	6
C	7	8	9

$$\text{Recall (A)} = \frac{TP}{TP+FN} = \frac{1}{1+2+3}$$

$$\text{Specificity (B)} = \frac{TN}{TN+FP} = \frac{1+9}{1+9+2+8}$$

$$\text{Precision (C)} = \frac{TP}{TP+FP} = \frac{9}{9+3+6}$$

CONCORDANT & DISCORDANT PAIRS

- **Discordant:** A pair of +ve and -ve observations where the model has no cutoff probability to correctly classify them
- **Concordant:** A pair of +ve and -ve observations where the model has a cutoff probability to correctly classify them
- If probability of +ve sample > probability of -ve sample, concordant pair

- If probability of +ve sample < probability of -ve sample, discordant pair

- Area under ROC curve is proportion of concordant pairs in the dataset

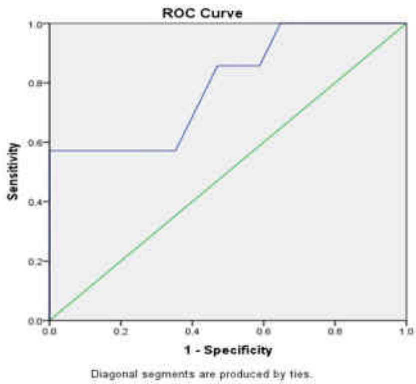
Hours of study	Passing exam	
	Probability	Label
1	0.070	0
2	0.260	0
3	0.610	0
4	0.870	1
5	0.950	1
6	0.970	1
7	0.980	0

(1,5): concordant pair
(4,7): discordant pair

ROC - Receiver Operating Characteristics

- ROC curve: FPR = 1-specificity = 1-TNR vs TPR = sensitivity (for diff threshold values)

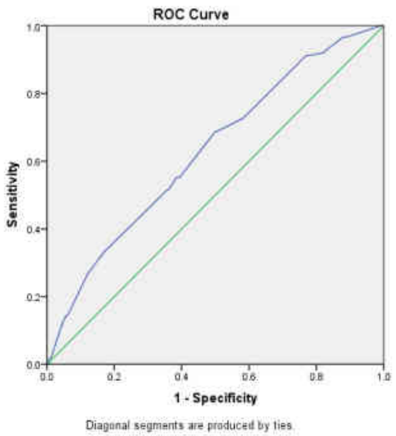
- Higher area ⇒ better prediction ability ⟶ check MI for details



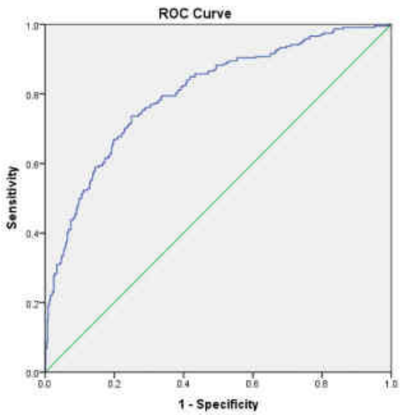
Area Under Curve (AUC)

- Probability that a model will rank a randomly chosen +ve higher than randomly chosen -ve

AUC = P(Random Positive Observation) > P(Random Negative Observation)



AUC = 0.629



AUC = 0.801

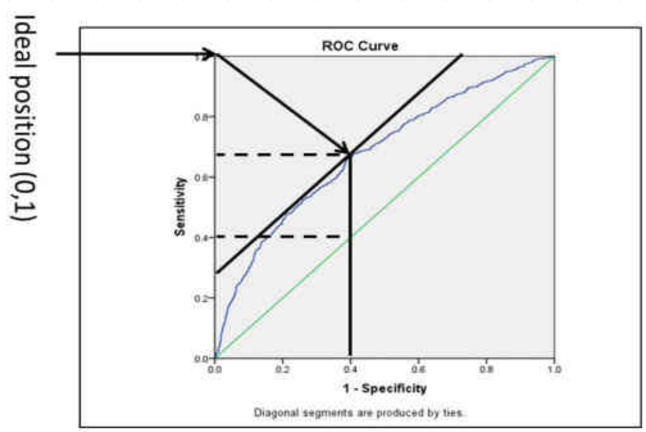
General Rule for Acceptance

- If area ≤ 0.5 ⟶ no discrimination
- If 0.7 ≤ area < 0.8 ⟶ acceptable
- If 0.8 ≤ area < 0.9 ⟶ excellent
- If area ≥ 0.9 ⟶ outstanding

Youden's Index for Optimal Cutoff Probability

- Best sensitivity for least FPR

Youden's Index = J statistic = $\max_{p} (\text{sensitivity}(p) + \text{specificity}(p) - 1)$



Cost-Based Cut-off Probability

- Penalize misclassifications — C_{01} : misclassify 0 as 1 (FP)
  C_{10} : misclassify 1 as 0 (FN)

Prediction

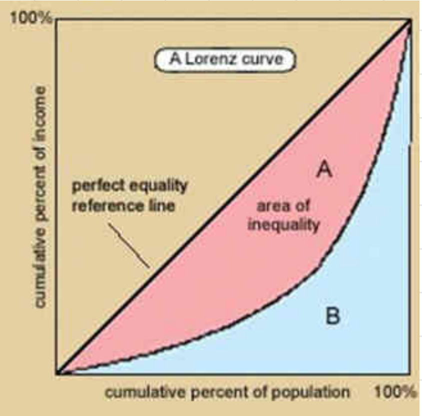
Actual	0	1
0	–	C_{01}
1	C_{10}	–

optimal cutoff probability

$$\min_{p} [C_{01} P_{01} + C_{10} P_{10}]$$

threshold

Lorenz curve



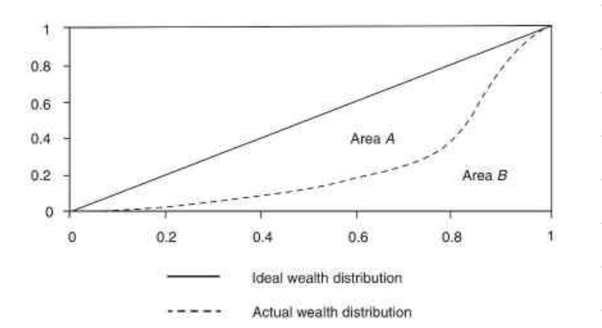


FIGURE 11.4 Lorenz curve.

Gini coefficient

- Different from Gini index used in decision trees
- Statistical measure of dispersion

$$\text{Gini coefficient} = \frac{A}{A+B} = 2\,AUC - 1$$

- https://stats.stackexchange.com/questions/342329/gini-and-auc-relationship

Q&A (CT1, pg 323)

1. In a multiple linear regression model, the overall model significance is tested using
   (a) Partial F-test (b) t-test (c) Durbin–Watson Test (d) F-test

Partial F-test: fit for partial model
t-test: significance of $\hat{\beta}_1$
DW: autocorrelation b/w successive error terms

2. In a multiple linear regression
   (a) R^2 and adjusted R^2 are non-decreasing functions.
   (b) R^2 is an increasing function, whereas adjusted R^2 is non-decreasing function.
   (c) R^2 and adjusted R^2 are increasing functions.
   (d) R^2 is a non-decreasing function and adjusted R^2 may increase or decrease.

3. Which of the following equation(s) is/are not multiple linear regression?
   (a) $\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
   (b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \frac{1}{1+\beta_3} X_1 X_2$
   (c) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
   (d) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 X_2^2$

4. In multiple regression models, multi-collinearity may result in
   (a) Removing a statistically significant explanatory variable from the model.
   (b) The regression coefficient may have opposite sign.
   (c) Adding a new variable to the model may cause huge change to the regression coefficient.
   (d) All of above.

Multicolinearity destabilizes model

5. When a new variable is added to the regression model, the R-square value increases by
   (a) Square of the semi-partial correlation between added variable and the response variable.
   (b) Correlation coefficient between the added variable and the response variable.
   (c) Partial correlation coefficient between added variable and the response variable.
   (d) Semi-partial coefficient between added variable and the response variable.

Look: VIF

6. If there is an auto-correlation between the successive errors in a time series regression then
   (a) A statistically insignificant variable may be added to the model.
   (b) A statistically significant variable may be removed from the model.
   (c) The standard error of estimate of the regression parameter is underestimate(d)
   (d) The Durbin–Watson test statistic value will be close to 2.

7. When a stepwise regression model is developed, the first variable that is added is
   (a) The variable with highest variance.
   (b) The variable that has the least variance.
   (c) The variable that has highest correlation with the dependent variable.
   (d) The variable with least covariance with the dependent variable.

8. Variance inflation factor is
   (a) Factor by which the regression coefficient is increase(d)
   (b) Factor by which the t-statistic value is inflate(d)
   (c) Factor by which the t-statistic is deflated by a factor of $\sqrt{VIF}$.
   (d) Factor by which the t-statistic value is inflated by a factor of $\sqrt{VIF}$.

9. Variable selection in stepwise regression is achieved through:
   (a) Partial F-test (b) F-test (c) Correlation (d) t-test

10. A regression model is developed between salary earned by a graduating MBA student using a sample of 450 students and their undergraduate discipline (where the base category is discipline "arts"). The regression output is shown in Table 10.39.

TABLE 10.39 Regression coefficients

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	198246.40	45690.10	4.338	1.8×10^{-05}
	Science	39430.00	20020.60	2.121	0.036
	Engineering	56940.50	22450.67	2.536	0.011
	Commerce	−14250.89	8932.45	1.5954	0.111

Which of the following statements are true at 5% significance:
   (a) Students from arts category earn minimum average salary.
   (b) Students from engineering category earn the maximum average salary.
   (c) The average salaries earned by arts and commerce graduates are same.
   (d) Science students earn 39430 more than arts students on average.

11. A regression model is developed for salary of employees of a company using gender (G), work experience (WE) and the interaction variable $G \times WE$. $G = 1$ is coded as female and $G = 0$ is male. The corresponding regression equation is shown below (assume that all predictors are significant):

$$Y = 45{,}490.50 + 3000.900 \times G + 1497.89\, WE - 990.75\, G \times WE$$

Which of the following statements are true?
   (a) Average salary of female employees is higher than male employees
   (b) Female employees earn 3000.90 more than male employees
   (c) Increase in salary with work experience for male employees is higher than female employees.
   (d) In the long run, male employees earn more than female employees.

12. Which of the following measures are used for identifying influential observations in the data?

(a) Cook's distance (b) Mahalanobis distance

(c) Leverage value (d) All of above

13. Transformation of variables will be useful to solve the following problem(s) in MLR:

(a) Multi-collinearity

(b) Outliers

(c) Heteroscedasticity

(e) None of above

14. Regression model was developed on a time-series data, the value of Durbin–Watson statistic value is 0.2. Then

(a) There is a significant correlation between the independent variable and dependent variable.

(b) There is a positive auto-correlation between errors.

(c) There is a negative auto-correlation between errors.

(d) There is no auto-correlation.

15. The independent variable that has the highest impact on the dependent variable is given by

(a) The variable with largest coefficient value.

(b) The variable with largest absolute coefficient value.

(c) The variable with largest standardized coefficient value.

(d) The variable with largest absolute standardized coefficient value.