Slides Prepared By
Dr. Hussein Hazimeh

Lebanese University
Faculty of Information 1

Data Science Departement

2nd year – Data Analysis in R

June– 2022 – Chapter 8

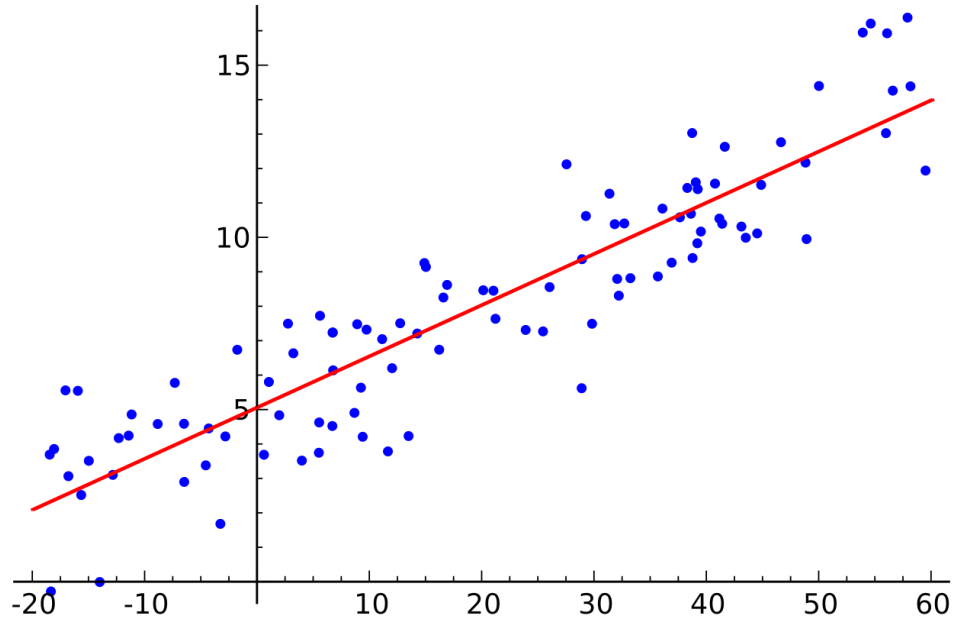
Agenda

- Time series analysis
- Moving average models
- Winters model
- Exponential smoothing techniques
- **Regression models**
- Measures for godness of the forecast
- Handling missing values in time series

Prediction by Regression

Predictions by Regression

- **Linear Regression**
- Linear regression is one of the simplest and most common supervised machine learning algorithms that data scientists use for predictive modeling.
- Statistical researchers often use a linear relationship to predict the (average) numerical value of *Y* for a given value of *X* using a straight line (called the *regression line*). If you know the slope and the *y*-intercept of that regression line, then you can plug in a value for *X* and predict the average value for *Y*. In other words, you predict (the average) *Y* from *X*.
- The scatterplot must form a linear pattern.
- More data leads to more better regression results.



Linear Regression

» With one independent variable, we may write the regression equation as:

$$Y = a + bX + e$$

» Where Y is an observed score on the dependent variable, *a* is the intercept, *b* is the slope, X is the observed score on the independent variable, and *e* is an error or residual.

Linear Regression

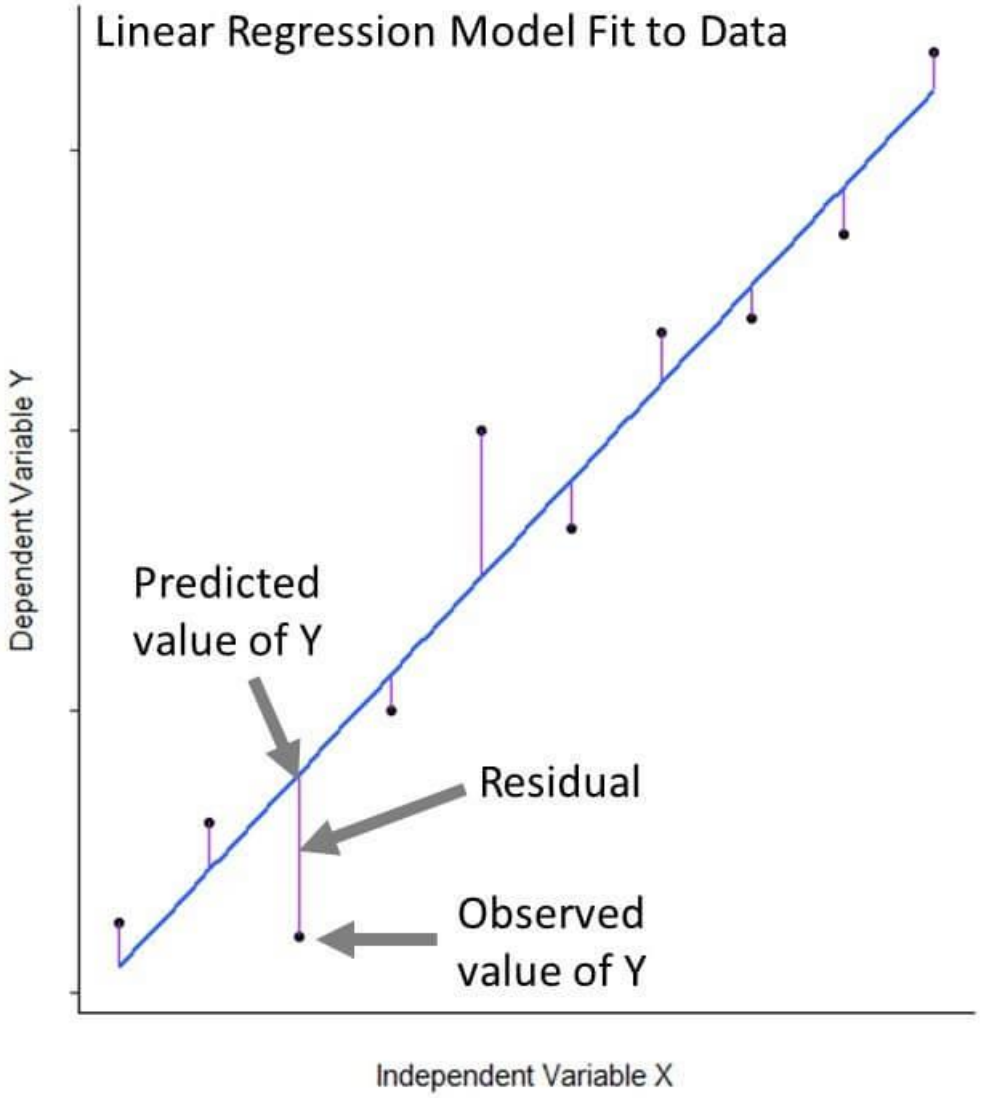
- A taxicab company manager believes that the monthly repair costs (Y) of cabs are related to age (X) of the cabs. Five cabs are selected randomly and from their records we obtained the following data: (x, y) = {(2, 2), (3, 5), (4, 7), (5, 10), (6, 11)}. Based on our practical knowledge and the scattered diagram of the data, we hypothesize a linear relationship between predictor X, and the cost Y.

- Now the question is how we can best (i.e., least square) use the sample information to estimate the unknown slope (m) and the intercept (b)? The first step in finding the least square line is to construct a sum of squares table to find the sums of x values (Σx), y values (Σy), the squares of the x values (Σx^2), the squares of the x values (Σy^2), and the cross-product of the corresponding x and y values (Σxy), as shown in the following table:

Linear Regression

Formulas and Notations:

- $\bar{x} = \Sigma x / n$
  This is just the mean of the x values.
- $\bar{y} = \Sigma y / n$
  This is just the mean of the y values.
- $S_{xx} = SS_{xx} = \Sigma(x(i) - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2 / n$
- $S_{yy} = SS_{yy} = \Sigma(y(i) - \bar{y})^2 = \Sigma y^2 - (\Sigma y)^2 / n$
- $S_{xy} = SS_{xy} = \Sigma(x(i) - \bar{x})(y(i) - \bar{y}) = \Sigma x \cdot y - (\Sigma x) \cdot (\Sigma y) / n$
- Slope $m = SS_{xy} / SS_{xx}$
- Intercept, $b = \bar{y} - m \cdot \bar{x}$
- y-predicted = $yhat(i) = m \cdot x(i) + b$.
- Residual(i) = Error(i) = $y - yhat(i)$.
- $SSE = S_{res} = SS_{res} = SS_{errors} = \Sigma[y(i) - yhat(i)]^2$.
- Standard deviation of residuals = $s = S_{res} = S_{errors} = [SS_{res} / (n-2)]^{1/2}$.
- Standard error of the slope (m) = $S_{res} / SS_{xx}^{1/2}$.
- Standard error of the intercept (b) = $S_{res}[(SS_{xx} + n. \bar{x}^2) / (n \cdot SS_{xx}]^{1/2}$.



Linear Regression

» Table:

	x	y	x^2	xy	y^2
	2	2	4	4	4
	3	5	9	15	25
	4	7	16	28	49
	5	10	25	50	100
	6	11	36	66	121
SUM	**20**	**35**	**90**	**163**	**299**

» The second step is to substitute the values of Σx, Σy, Σx^2, Σxy, and Σy^2 into the following formulas:

» $SS_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)/n = 163 - (20)(35)/5 = 163 - 140 = 23$

» $SS_{xx} = \Sigma x^2 - (\Sigma x)^2/n = 90 - (20)^2/5 = 90 - 80 = 10$

» $SS_{yy} = \Sigma y^2 - (\Sigma y)^2/n = 299 - 245 = 54$

Linear Regression

» Table:

	x	y	x^2	xy	y^2
	2	2	4	4	4
	3	5	9	15	25
	4	7	16	28	49
	5	10	25	50	100
	6	11	36	66	121
SUM	**20**	**35**	**90**	**163**	**299**

» Use the first two values to compute the estimated slope:

» Slope = m = SS_{xy} / SS_{xx} = 23 / 10 = 2.3

Linear Regression

» Table:

	x	y	x^2	xy	y^2
	2	2	4	4	4
	3	5	9	15	25
	4	7	16	28	49
	5	10	25	50	100
	6	11	36	66	121
SUM	**20**	**35**	**90**	**163**	**299**

» To estimate the intercept of the least square line, use the fact that the graph of the least square line always pass through (,) point, therefore,

» The intercept = b = $\bar{y}$ – (m)($\bar{x}$) = (Σy)/ 5 – (2.3) (Σx/5) = 35/5 – (2.3)(20/5) = -2.2

» Therefore the least square line is:

y-predicted = yhat = mx + b = -2.2 + 2.3x

Linear Regression

» Table:

x Predictor	-2.2+2.3x y-predicted	y observed
2	2.4	2
3	4.7	5
4	7	7
5	9.3	10
6	11.6	11

y-predicted = yhat = mx + b = -2.2 + 2.3x

Linear Regression

» Table:

x Predictor	-2.2+2.3x y-predicted	y observed
2	2.4	2
3	4.7	5
4	7	7
5	9.3	10
6	11.6	11

» After estimating the slope and the intercept the question is how we determine statistically if the model is good enough, say for prediction. The standard error of slope is:

» Standard error of the slope (m)= $S_m = S_{res} / S_{xx}^{1/2}$, $t_{slope} = m / S_m$,For our numerical example, it is:

$$t_{slope} = 2.3 / [(0.6055)/ (10^{1/2})] = 12.01$$

which is large enough, indication that the fitted model is a "good" one.

Linear Regression

» You may ask, in what sense is the least squares line the "best-fitting" straight line to 5 data points. The least squares criterion chooses the line that minimizes the sum of square vertical deviations, i.e., residual = error = y - yhat:

$$SSE = \Sigma (y - yhat)^2 = \Sigma (error)^2 = 1.1$$

» The numerical value of SSE is obtained from the following computational table for our numerical example

x Predictor	-2.2+2.3x y-predicted	y observed	error y	squared errors
2	2.4	2	-0.4	0.16
3	4.7	5	0.3	0.09
4	7	7	0	0
5	9.3	10	0.7	0.49
6	11.6	11	-0.6	0.36
			Sum=0	**Sum=1.1**

» Alternately, one may compute SSE by: $SSE = SS_{yy} - m\, SS_{xy} = 54 - (2.3)(23) = 54 - 52.9 = 1.1$

Further reading

- http://home.ubalt.edu/ntsbarsh/Business-stat/stat-data/Forecast.htm#rhowma