An Assignment on Linear Regression

Course Name: Applied Regression Analysis

Course Code: PM-ASDS08

Submitted to:

Sultana Begum

Assistant Professor

Department of Statistics, JU

Submitted by:

Mohammad Saiduzzaman Sayed

ID: 20215063

Batch: 5th

Sec: A

Professional Masters in

Applied Statistics and Data Science (PM-ASDS)

JAHANGIRNAGAR UNIVERSITY

Practical 1

Database Description:

Description:

To predict the sales based on the impact of three advertising media which are YouTube, Facebook, and Newspaper. Our dataset has 200 observations and 4 variables.

Format:

The variables are described as

y: Sales

x_1: YouTube advertising budget

x_2: spending on Facebook advertisements

x_3: advertising cost in newspaper

Note: figures are in thousands of dollars

Descriptive Statistics:

	y	x_1	x_2	x_3
Min	1.92	0.84	0.00	0.36
1st Qu.	12.45	89.25	11.97	15.30
Median	15.48	179.70	27.48	30.90
Mean	16.83	176.45	27.92	36.66
3rd qu.	20.88	262.59	43.83	54.12
Max	32.40	355.68	59.52	136.80

Assumptions:

~ The relationship between dependent variable and independent variable is linear.

~ Errors are normally distributed (with mean zero and constant variance).

~ Errors are homoscedastic.

~ Observations are independent.

~ Regressors (input variables) are not correlated.

(a)

(i) Verify the normality of response variable:

Histogram:

Histograms provide a graphical display to help us assess normality. In the histogram below, we can say that our Sales data is approximately normally distributed.

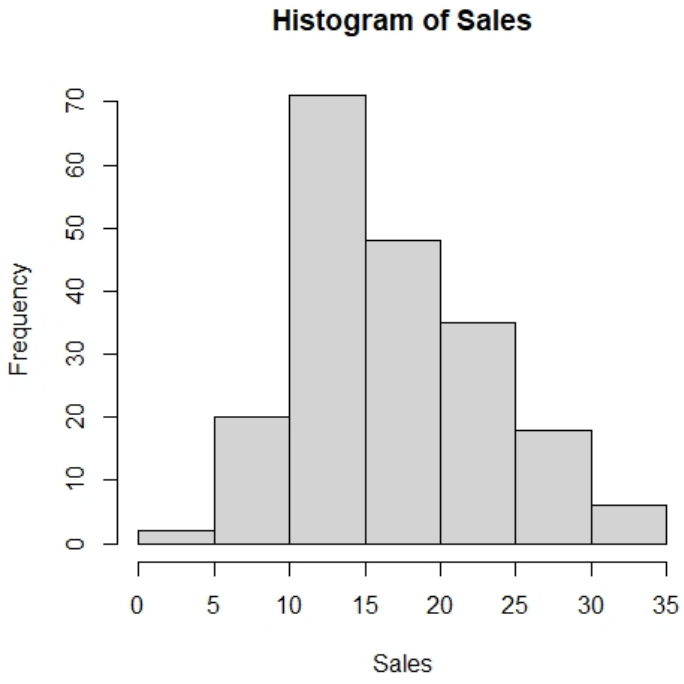

Fig: Histogram of Sales

Normal Q-Q plot:

Another graph to assess normality is the Q-Q plot. In the univariate Normality test using Q-Q plot, If the line is 45-degree straight, then we can say our data is normally distributed. In this case the plot below, the straight line is near the 45-degree, which means our Sales data is approximately normally distributed.

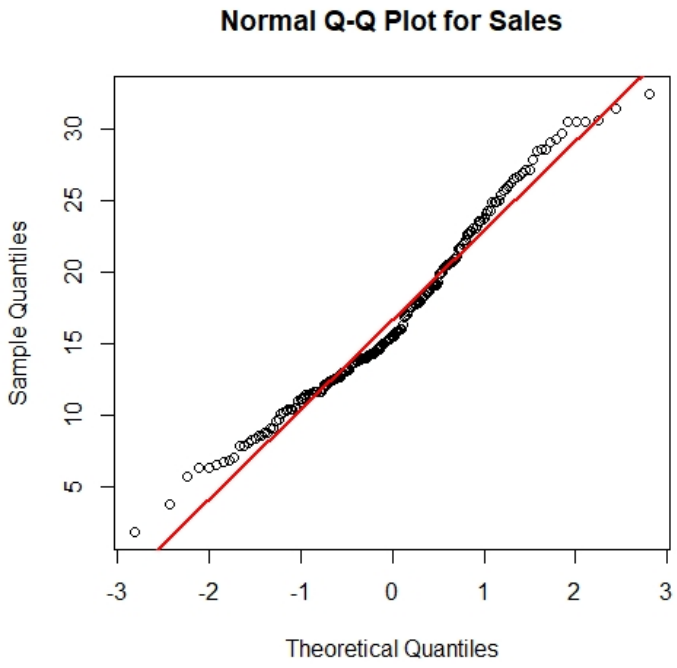

Fig: Normal Q-Q plot of Sales

(ii) Linear association between Sales and three advertising costs:

Scatter Plot and Correlation Matrix:

We can see the relationship between Sales and three advertising costs from the below scatter plot. And correlation between Sales and three advertising costs.

Sales(y) and YouTube advertising budget (x_1) ~ Moderate Positive Correlation (78%)

Sales(y) and spending on Facebook advertisements (x_2) ~ Moderate Positive Correlation (58%)

Sales(y) and advertising cost in newspaper (x_3) ~ Weak Positive Correlation (23%).

Multicollinearity Checking:

YouTube advertising budget (x_1) and Facebook advertisements (x_2) ~ Moderate Positive Correlation (55%)

YouTube advertising budget (x_1) and advertising cost in newspaper (x_3) ~ Moderate Positive Correlation (57%)

Facebook advertisements (x_2) and advertising cost in newspaper (x_3) ~ weak Positive Correlation (35%)

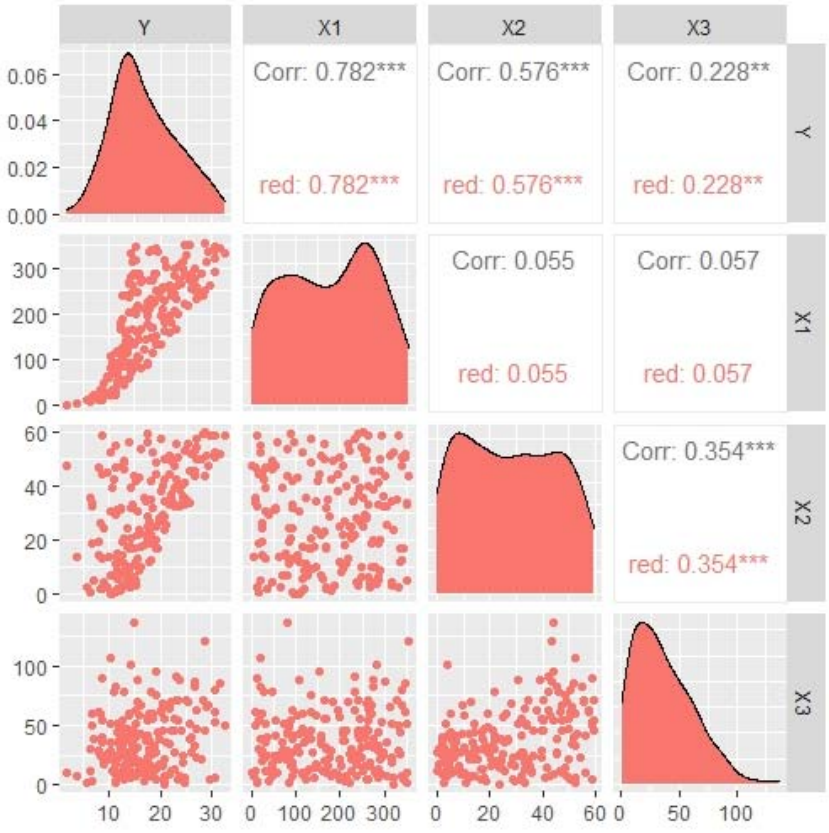

Fig: Scatter plot and Correlation matrix

(iii) Test of Significance of Regression:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_a: At\ least\ one\ \beta_i \neq 0\ \ i = 1,2,3$$

Test Statistic, F = $\frac{MSR}{MSE}$ = 570.3

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	3.526667	0.374290	9.422	<2e-16	***
x1	0.045765	0.001395	32.809	<2e-16	***
x2	0.188530	0.008611	21.893	<2e-16	***
x3	-0.001037	0.005871	-0.177	0.86	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Decision:

Reject the null hypothesis at 5% α level of significance if p-value < α. Here, we reject null hypothesis at 95% confidence level because of p-value < 0.05. And we accept the alternative hypothesis which means, it is reasonable to apply regression analysis here.

R^2_{adj}:

Adjusted R-square means proportion of total sample variability of y explained by the linear relationship between response and regressors.

$R^2_{adj} = 1 - \frac{SSE}{SST}$ = 0.8956 which means about 89.54% 0f sample variation in Sales is explained by $\hat{y} = 3.51 + 0.046\ x_1 + 0.19\ x_2 - 0.001\ x_3$

(iv) Test of Significance of Individual Regressor:

For YouTube advertising budget,

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0 \ (\beta_1 > 0 \ or \ \beta_1 < 0)$$

Test statistic $t = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)} = 32.809$

Decision:

Here, we reject null hypothesis at 5% significant level because of p-value < 0.05. And we accept the alternative hypothesis which means, YouTube advertising budget is significant for Sales.

For spending on Facebook advertisements,

$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 \neq 0 \ (\beta_2 > 0 \ or \ \beta_2 < 0)$$

Test statistic $t = \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} = 21.893$

Decision:

Here, we reject null hypothesis at 5% significant level because of p-value < 0.05. And we accept the alternative hypothesis which means, spending on Facebook advertisements is significant for Sales.

For advertising cost in newspaper,

$$H_0: \beta_3 = 0$$

$$H_a: \beta_3 \neq 0 \ (\beta_3 > 0 \ or \ \beta_3 < 0)$$

Test statistic $t = \frac{\widehat{\beta}_3}{se(\widehat{\beta}_3)} = -0.177$

Decision:

Here, we cannot reject null hypothesis at 5% significant level because of p-value > 0.05. And we cannot accept the alternative hypothesis which means, advertising cost in newspaper is not significant for Sales.

(v) The best fitted model:

So, advertising cost in newspaper is not significant for Sales (x_3), we can fit our model without x_3 variable. Here, our best fitted model is,

$$\hat{y} = 3.51 + 0.046\, x_1 + 0.19\, x_2$$

And Std. Error (0.001395) (0.008611)

$$R^2_{adj} = 0.8956$$

Interpretation:

1. The intercept coefficient is 3.51, it means that for advertising budget on both medias equal to zero, we can expect sales of 3.51 *1000 = $3,510
2. The regression coefficient (β1) shows that if we increase (decrease) YouTube advertising budget equal to 1000 dollars, we can expect sales to increase (decrease) by an average of $46 (0.046*1000), controlling other variables.
3. The regression coefficient (β2) indicates that if we increase (decrease) Facebook advertising budget equal to 1000 dollars, we can expect sales to increase (decrease) by an average of $190 (0.19*1000).

(vi) Estimate mean sales of all products where $x_1 = 50, x_2 = 40 \; and \; x_3 = 20$.

Our model is,

$$\hat{y} = 3.51 + 0.046\, x_1 + 0.19\, x_2 - 0.001\, x_3$$

$$= 3.51 + 0.0460 * (50) + 0.19 * (40) - 0.001(20)$$

$$= 13.39$$

We can say that, if our YouTube advertising budget is \$50000, spending on Facebook advertisements \$40000, and advertising cost in newspapers \$20000 then we can expect sales of \$13390 $(13.39 * 1000)$.

(vii) Diagnostic plot:

Residuals vs Fitted Values Plot:

A Residuals vs Fitted plot displays the spread of residuals over fitted values. If the red line across the center of the plot is roughly horizontal, then we can assume that the residuals follow a linear pattern (linearity). From below plot, we say that the residuals follow approximately a linear pattern.

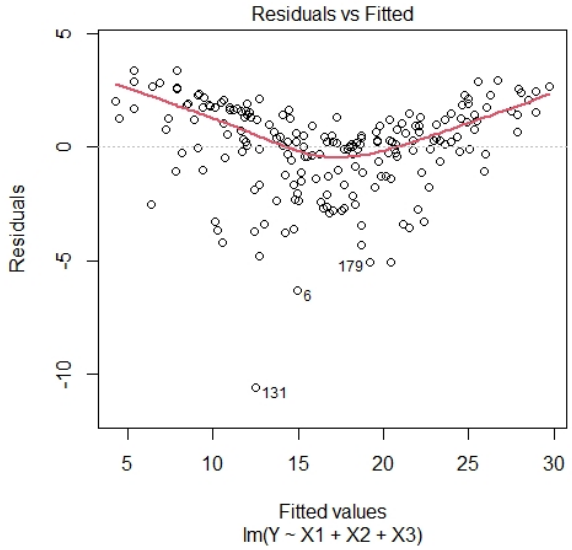

Fig: Residuals vs Fitted Values

Normal Q-Q plot of Residuals:

We can see that the points fall roughly along the straight diagonal line although the observation #131 deviates a bit from the line at the tail end in the below plot. So, the residuals are approximately normally distributed.

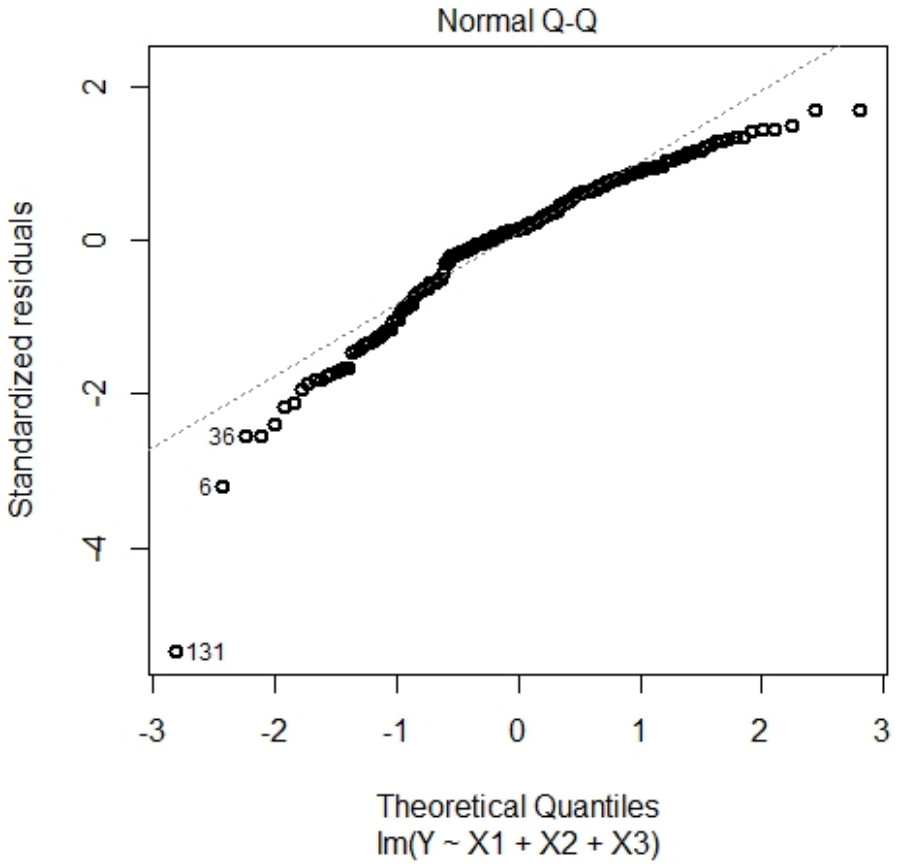

Fig: Normal Q-Q plot

Scale-Location plot:

A scale-location plot displays the fitted values of a regression model. By this plot, we can check whether the red line is approximately horizontal or not. And we can check whether the spread around the red line with the fitted values does vary or not. We can see that in below plot, the red line is approximately horizontal across the plot. The assumption of equal variance is not violated. So, we can say that Errors are homoscedastic.

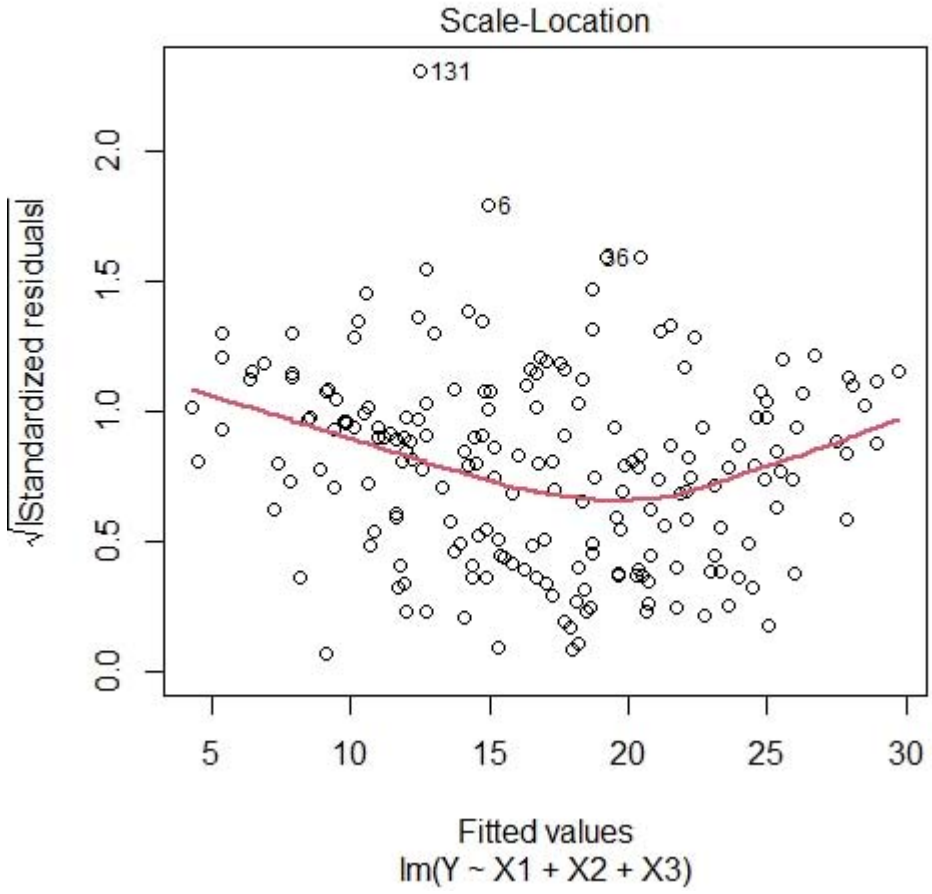

Fig: Scale-Location plot

Residuals vs Leverage plot:

A Residuals vs Leverage plot represents to identify influential observations in a regression model. If any point in the plot falls outside of Cook's distance (the red dashed lines) then it is considered to an influential observation. In the above plot, we can see that observation #131 lies closest to the border of Cook's distance, but it doesn't fall outside of the dashed line. This means there are not any influential points in our regression model.

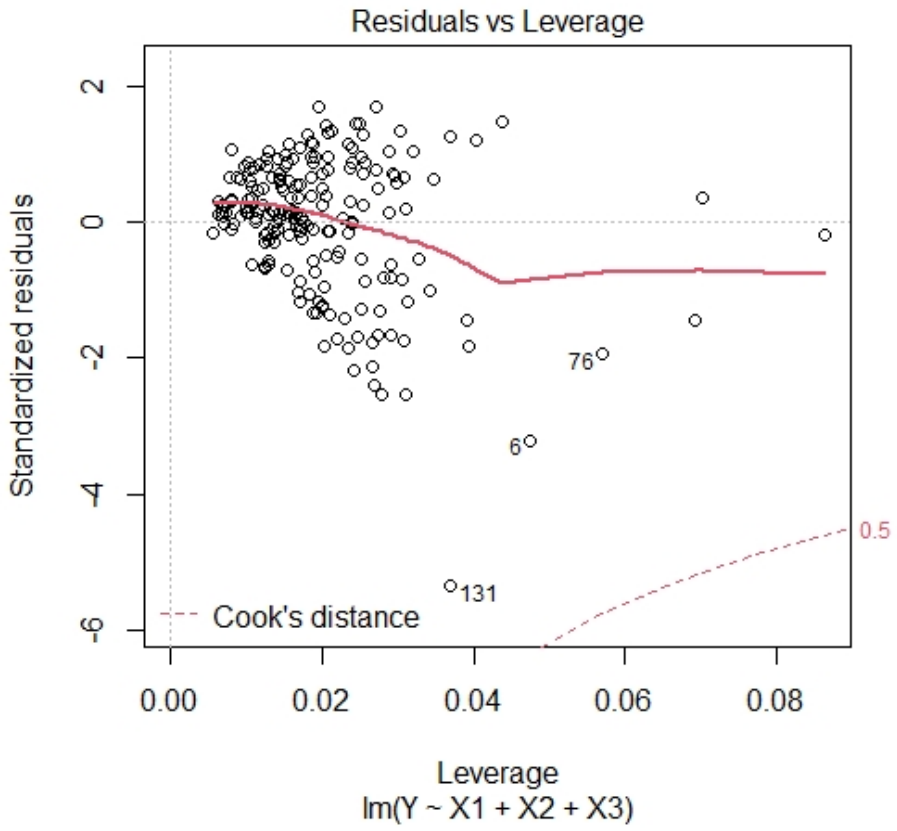

Fig: Residuals vs Leverage plot

(b) Variable Selection:

~ Stepwise regression (Forward or Backward Method)

~ All possible regression: 2^m -1 regression models for m predictors

Variable Selection Criteria:

~ Adjusted R^2: High value is preferable

~ Mallow's C_p**:** Low value is preferable

~ AIC (Akaike Information Criteria): Low value is preferable

~ BIC (Bayesian Information Criteria): Low value is preferable

All possible regression:

Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp	AIC	BIC
1	1	x1	0.61	0.60	544.08	1117.0200	545.4604
2	1	x2	0.33	0.32	1077.68	1225.6024	653.2339
3	1	x3	0.05	0.04	1611.42	1295.6000	722.8853
4	2	x1, x2	0.89	0.89	2.03	853.3227	285.8680
5	2	x1, x3	0.64	0.64	481.32	1100.7068	527.9175
6	2	x2, x3	0.33	0.32	1078.40	1227.4009	653.3479
7	3	x1, x2, x3	0.89	0.89	4.00	855.2909	287.8779

Since there is m = 3 regressors, there are 2^3 -1 = 7 possible regression equations. The results of fitting these 7 equations are displayed in above table. Here, 4 no. the equation has the highest R-Square (0.89) and Adj. R-Square (0.89). Also, it has the lowest value of Mallow's Cp (2.03), AIC (853.3227), and BIC (285.8680).

So, we can say that x1 and x2 both are significant regressors and the fitted model with x1 and x2 is the best-fitted model.

Stepwise Regression:

Model Index	Predictors	R-Square	Adj. R-Square	Pred R-Square	Mallow's Cp	AIC	BIC
1	x1	0.61	0.6099	0.6034	544.0814	1117.0200	545.4604
2	x1, x2	0.89	0.8962	0.8925	2.0312	853.3227	285.8680
3	x1, x2, x3	0.89	0.8956	0.8912	4	855.2909	287.8779

In stepwise regression we also see that, x1 and x2 both are significant regressors and the fitted model with x1 and x2 is the best-fitted model because of highest R-Square, Adj. R-Square and Pred R-Square where lowest value of Mallow's Cp, AIC, and BIC.

Conclusion:

After verifying our all assumption now our model is ready for apply regression analysis. And here x1 (YouTube advertising budget) and x2 (spending on Facebook advertisements) variables are significant for y (Sales). So, our best fitted linear regression equation is,

$$\hat{y} = 3.51 + 0.046\, x_1 + 0.19\, x_2$$

THE END