Lecture

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: SY BSc

Subject : Linear regression in R

Subject Code: SMR

Chapter: Unit 2 Chp 1

Chapter Name: Introduction to Linear regression

R-square

INTERPRETATION OF R^2

- IF $R^2 = 1$, ALL OF THE DATA POINTS FALL PERFECTLY ON THE REGRESSION LINE. THE PREDICTOR *X* ACCOUNTS FOR *ALL* OF THE VARIATION IN *Y*!
- IF $R^2 = 0$, THE ESTIMATED REGRESSION LINE IS PERFECTLY HORIZONTAL. THE PREDICTOR *X* ACCOUNTS FOR *NONE* OF THE VARIATION IN *Y*!
- IF R^2 IS BETWEEN 0 AND 1, "$R^2 \times 100$ PERCENT OF THE VARIATION IN *Y* IS 'EXPLAINED BY' THE VARIATION IN PREDICTOR *X*."

Interpretation of R-square

- In the context of predictive models (usually linear regression), where *y* is the true outcome, and *f* is the model's prediction, the definition that I see most often is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- In words, R^2 is a measure of how much of the variance in *y* is explained by the model, *f*.
- Or
- R^2 IS ALSO THE SQUARE OF THE CORRELATION (CORRELATION WRITTEN AS A "P" OR "RHO") BETWEEN THE ACTUAL AND PREDICTED OUTCOMES.

Using p Value To Check For Statistical Significance

1. The summary statistics above tells us a number of things.
2. One of them is the model's p-Value (in last line) and the p-Value of individual predictor variables
3. The p-Values are very important.
4. Because, we can consider a linear model to be statistically significant only when both these p-Values are less than the pre-determined statistical significance level of 0.05.
5. This can visually interpreted by the significance stars at the end of the row against each X variable.
6. The more the stars beside the variable p-Value, the more significant the variable.

What is the Null and Alternate Hypothesis?

1. Whenever there is a p-value, there is always a Null and Alternate Hypothesis associated.
2. So what is the null hypothesis in this case?
3. In Linear Regression, the Null Hypothesis (H0) is that the beta coefficients associated with the variables is equal to zero.
4. The alternate hypothesis (H1) is that the coefficients are not equal to zero. (i.e. there exists a relationship between the independent variable in question and the dependent variable).

What is t-value?

1. We can interpret the t-value something like this. A larger t-value indicates that it is less likely that the coefficient is not equal to zero purely by chance.
2. So, higher the t-value, the better.
3. Pr(>|t|) or p-value is the probability that you get a t-value as high or higher than the observed value when the Null Hypothesis (the ? coefficient is equal to zero or that there is no relationship) is true.
4. So if the Pr(>|t|) is low, the coefficients are significant (significantly different from zero).
5. If the Pr(>|t|) is high, the coefficients are not significant.

Standard Error and F-Statistic

Both standard errors and F-statistic are measures of goodness of fit.

$$Std.\,Error = \sqrt{MSE} = \sqrt{\frac{SSE}{n-q}}$$

$$F-statistic = \frac{MSR}{MSE}$$

where, n is the number of observations, q is the number of coefficients and MSR is the mean square regression, calculated as,

$$MSR = \frac{\sum_i^n \left(\hat{y_i} - \bar{y}\right)}{q-1} = \frac{SST - SSE}{q-1}$$

The higher the F-Statistic the better it is.

MSE: Mean squared error
MSR: Mean square regression

SSE: Sum of squares error

SST: Sum of squares total

SSR: Sum of square regression

Calculation of f-statistic in R

```
MSR=sum((fitted(model1)-mean(df_tr$sales))**2)/3
MSR
MSE=sum((residuals(model1))**2)/(150-4)
MSE

f_value=MSR/MSE
f_value
```

```
> summary(model1)$fstatistic
   value    numdf    dendf
443.1946   3.0000 146.0000
> MSR=sum((fitted(model1)-mean(df_tr$sales))**2)/3
> MSR
[1] 1882.795
> MSE=sum((residuals(model1))**2)/(150-4)
> MSE
[1] 4.248236
>
> f_value=MSR/MSE
> f_value
[1] 443.1946
```

What is the SSR?

The second term is the sum of squares due to regression, or SSR. It is the sum of the differences between the *predicted* value and the mean of the *dependent variable*. Think of it as a measure that describes how well our line fits the data.

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

What is the SST?

The sum of squares total, denoted SST, is the squared differences between the observed *dependent variable* and its mean.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2$$

What is the SSE?

1. **The last term is the sum of squares error, or SSE.**
2. **The error is the difference between the *observed* value and the *predicted* value.**

How Are They Related?

Mathematically, SST = SSR + SSE.

1. Total variability = Explained variability + Unxplained variability

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2$$

What is AIC and BIC?

1. The Akaikes information criterion – AIC (Akaike, 1974)
2. and the Bayesian information criterion – BIC (Schwarz, 1978)
3. are measures of the goodness of fit of the linear regression model
4. and can also be used for model selection.
5. where, n is the sample size.
6. For model comparison, the model with the lowest AIC and BIC score is preferred.

AIC = n* log(sum of squares error/n) + 2K
Where:

K is the number of model parameters (the number of variables in the model plus the intercept).

The lower the number, the better the fit.

The **Bayesian Information Criterion (BIC)** is almost the same as the AIC, although it tends to favor models with fewer parameters.

What is AIC and BIC?

1. The Akaikes information criterion – AIC (Akaike, 1974)
2. and the Bayesian information criterion – BIC (Schwarz, 1978)
3. are measures of the goodness of fit of the linear regression model
4. and can also be used for model selection.
5. where, n is the sample size.
6. For model comparison, the model with the lowest AIC and BIC score is preferred.

AIC = n* log(sum of squares error/n) + 2K
Where:

K is the number of model parameters (the number of variables in the model plus the intercept).

The lower the number, the better the fit.

The **Bayesian Information Criterion (BIC)** is almost the same as the AIC, although it tends to favor models with fewer parameters.

Multiple Linear Regression

- **Multiple linear regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the values of two or more other variables.**

- **The variable we want to predict is called as dependent variable (or sometimes response variable).**

- **The variables used to predict the value of dependent variable are called as independent variables (or sometimes, the predictor, explanatory or regressor variables).**

Statistical Model in Multiple Linear Regression

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \text{----} + b_p x_p + e$$

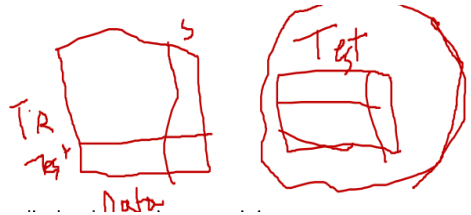
Where,

Y	: Dependent Variable
$x_1, x_2, \ldots, x_p$	: Independent Variables
$b_0, b_1, \ldots, b_p$	: Parameters of Model
e	: Random Error Component

Parameters of the model are estimated by Least Square Method.

Train-Test Split

- Before we build any predictive model, we should split the dataset into a *training* and a *test* data.
- The training dataset, as the name suggests, is used to learn the patterns from the data and build a model which gives us near-expected predictions.
- After we have built a *model* using the *training* data, we use the *test* data to estimate the performance of our model.
- Using *test* data gives us a completely unbiased estimate of our model accuracy and helps us understand how our model would perform in a real-world scenario.
- The splitting of the dataset into *training* and *test* should be random. Generally a 70:30 split is taken, i.e. 70% data as training and 30% data for testing.



Co-efficient of Determination

R_2

1. **R squared** is the proportion of variation in the response variable explained by the independent variables in the model.

$$R2 = \frac{explained\ variation}{\text{Total variation}}$$

- Both $R2$ and $R2$
- take value between 0 and 1, where 1 means 100% variation is explained

3. R2 increases or remains the same when independent variables are added to the model, even when the independent variables don't improve the fit.
   - Higher the value of $R2$ is our model.

Adjusted R2

1. **Adjusted R squared** as the name suggests is the proportion of variation explained by the model adjusted for the number of independent variables.
2. $R^2adj = 1 \frac{(1-R2)(n-1)}{n-p-1}$, for a model with n – observations and p independent variables.
3. R_{adj} increases only when the new variable improves the model fit more than expected by chance alone.

by the model. Due to the increasing nature of $R2$, it is not a good criterion to compare models.

Global Testing

F Test

A *F-test* is conducted to ascertain the whether the relationship between the dependent variable and the independent variables is statistically significant.

- The assumptions of the *F-test* are :

1. Error terms are normally distributed.
2. Error terms have mean 0 and common variance $\sigma 2$
3. Error terms are independent across observations

- The Hypothesis of the *F-test* is

$$H0: \beta i = 0 \; for \; all \; i \; vs \; H1: Atleast \; one \; \beta i \neq 0$$

- We reject $H0$ if the *p-value* of this test is less than 5% and conclude that our model is a good fit.
- This test is called the *Global Test of model adequacy*

Individual Testing

t-test

- The *t-test* checks whether a specific independent variable has a statistically significant impact on the dependent variable. The Hypothesis for this test is

$$H0: \beta i = 0 \; vs \; H1: \beta i \neq 0$$

- This t-*test* allow us to conclude whether each variable is statistically significant individually and therefore helps us in making the decision regarding whether to include that variable in our model.
- We reject $H0$ if *p-value* is less than 0.05 i.e., the variable has a significant impact on the dependent variable.
- We conduct this test for all the independent variables in our model separately.

Model Building

After identifying the variables influencing sales, we proceed with model building. We require to split the data for the purpose.

- 'caTools' has functions for random splitting of data.
- Using sample.split() we split the data into training and testing dataset.
- sample.split() gives logical values to the dataset and splits it into specified ratio for training and testing data.
- It also requires us to fixate on a field with respect to which it splits. Any field can be used.

```
#model building
library(caTools)
set.seed(101)
sample=sample.split(marketing$sales,SplitRatio = 0.7)
train=subset(marketing,sample==TRUE)
test=subset(marketing,sample==FALSE)
model=lm(sales~.,data=train)
summary(model)
```

Model selection

Primer

Primer

The model that includes all available explanatory variables is often referred to as the **full model**.

Model selection

P-values provide helpful information.

- Regression model relating price of a video game to various features, e.g. `cond_new` (1 if new, 0 if used), `stock_photo` (stock photo used).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
$R^2_{adj} = 0.7108$				$df = 136$

Model selection

Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called *backward-selection* and *forward-selection*.

Model selection

Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called *backward-selection* and *forward-selection*.

- The **backward-elimination** strategy starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until only variables with statistically significant p-values remain.
- The **forward-selection** strategy is the reverse of the backward-elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that present strong evidence of their importance in the model.

Backward elimination

- Start with the full model, and first eliminate `duration`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
$R^2_{adj} = 0.7108$				$df = 136$

Backward elimination

- To give the final model, after backward-selection.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.7849	0.7066	52.06	0.0000
cond_new	5.5848	0.9245	6.04	0.0000
wheels	7.2328	0.5419	13.35	0.0000
$R^2_{adj} = 0.7124$				$df = 138$

What is a Parsimonious Model?

Parsimonious models are simple models with great explanatory predictive power.

They explain data with a minimum number of parameters, or predictor variables.

The idea behind parsimonious models stems from “the law of briefness” (sometimes called *lex parsimoniae* in Latin).

The law states that you should use no more “things” than necessary;

In the case of parsimonious models, those “things” are parameters.

Parsimonious models have optimal parsimony, or just the right amount of predictors needed to explain the model well.

Parsimonius model

There is generally a tradeoff between goodness of fit and parsimony:

low parsimony models (i.e. models with many parameters) tend to have a better fit than high parsimony models.

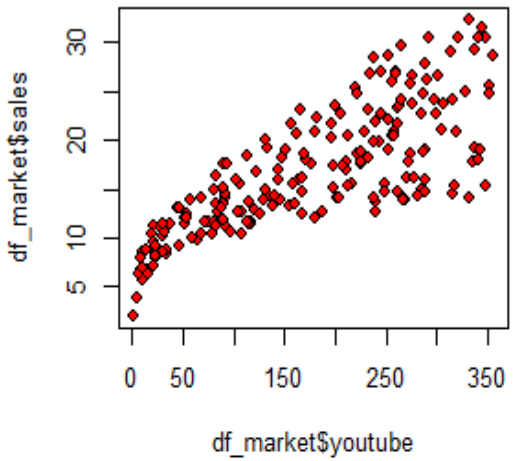
This is not usually a good thing;

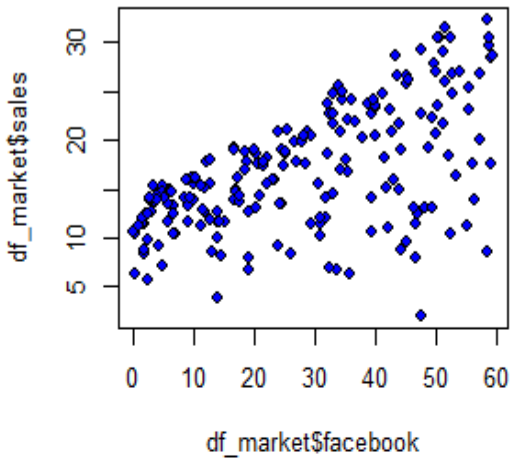
adding more parameters usually results in a good model fit for the data at hand,

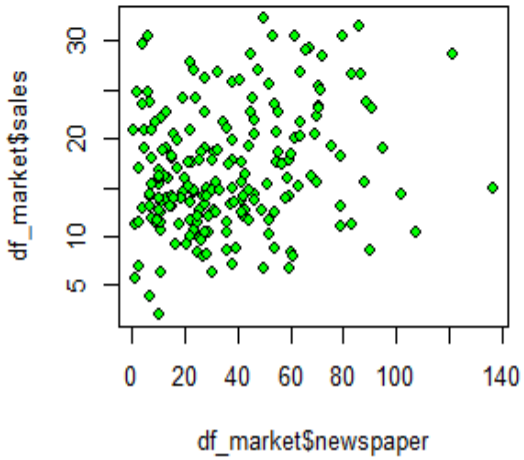
but that same model will likely be useless for predicting other data sets.

Linear relationship

- First, linear regression needs the relationship betw to be linear.
- It is also important to check for outliers since linea
- The linearity assumption can best be tested with



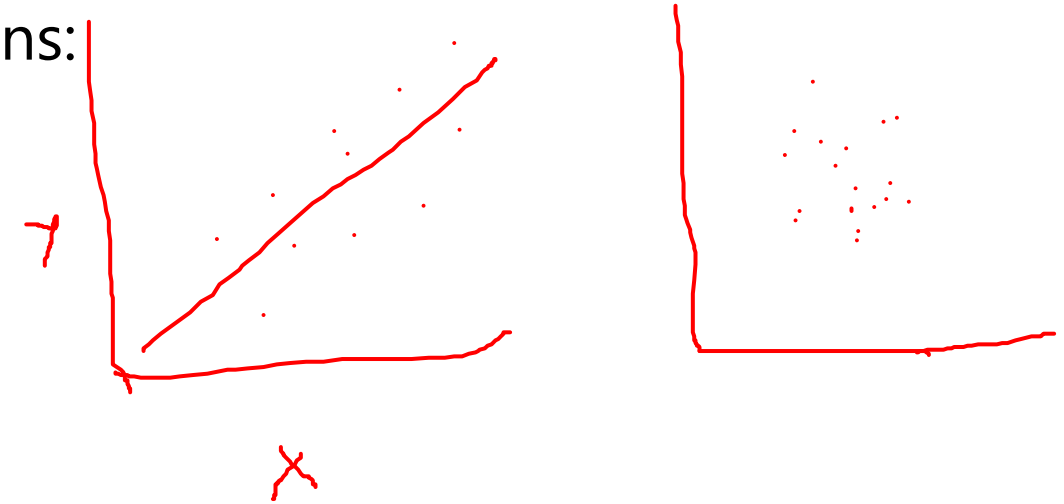





```
par(mfrow=c(2,2))

plot(x=df_market$youtube,y=df_market$sales,pch=21,cex=1,bg="red")

plot(x=df_market$facebook,y=df_market$sales,pch=21,cex=1,bg="blue")

plot(x=df_market$newspaper,y=df_market$sales,pch=21,cex=1,bg="green"
)
```

Assumptions of Linear Regression

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:



Linear relationship

Multivariate normality

No or little multicollinearity

No auto-correlation

Homoscedasticity

Multivariate normality

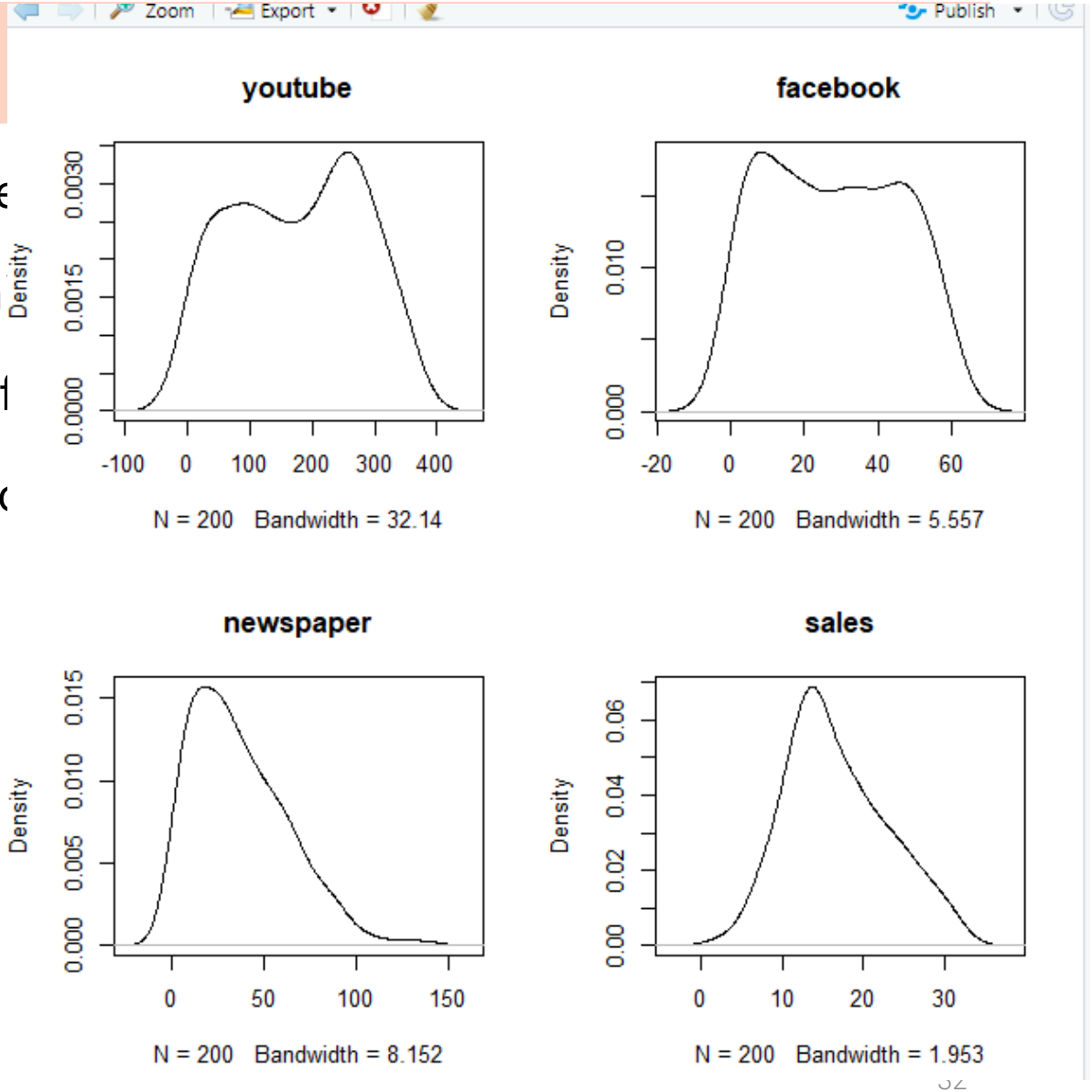
Secondly, the linear regression analysis require

This assumption can best be checked with a h

Normality can be checked with a goodness of

When the data is not normally distributed a no transformation) might fix this issue.

```
dev.off()
par(mfrow=c(2,2))
plot(density(x=df_market$youtube),main="youtube")
plot(density(x=df_market$facebook),main="facebook")
plot(density(x=df_market$newspaper),main="newspaper")
plot(density(x=df_market$sales),main="sales")
```



Log transformation using R

Sometimes, we have to deal with data that don´t have a normal shape but a skewed one.

A negative skewness reveals that the mean of the values is less than the median,

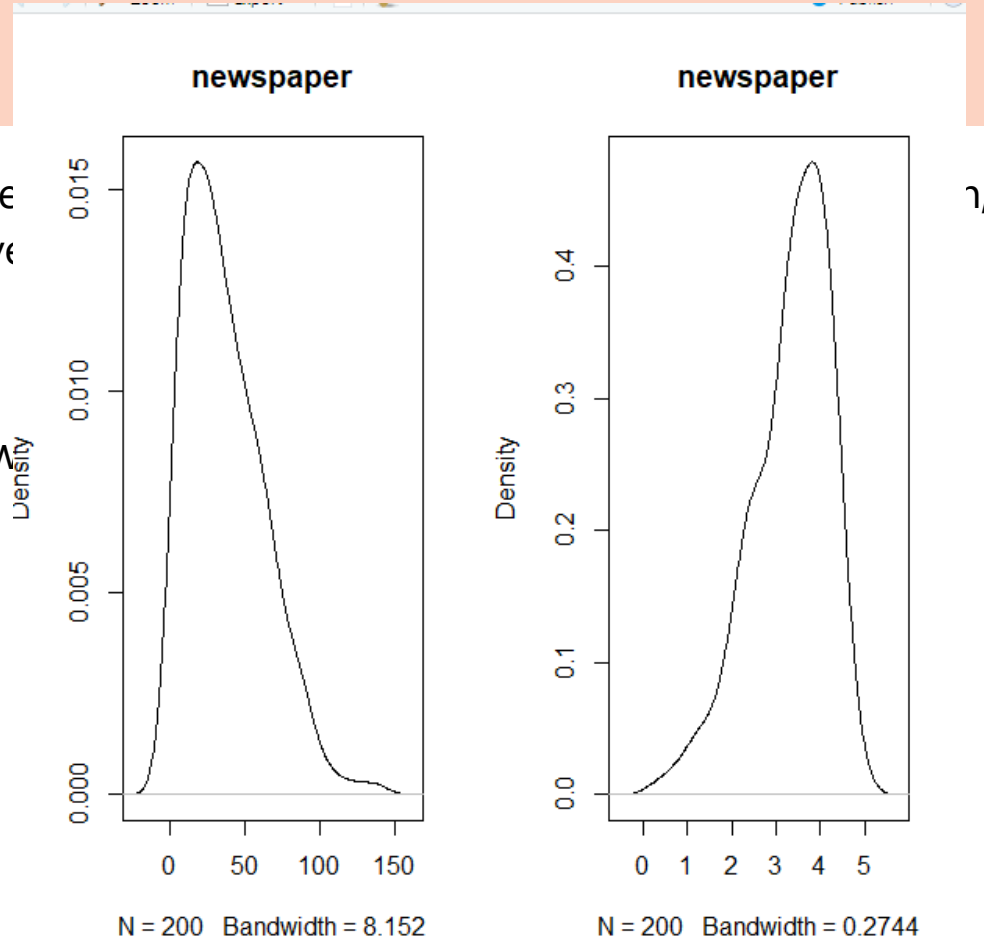
which means that the data distribution is left-skewed.

A positive skewness suggests that the mean of the data values is larger than the median,

and the data distribution is right-skewed.

log-transformation in the values might help us to improve
which, by default, computes the natural logarithm of a give

```
> par(mfrow=c(1,2))

> plot(density(x=df_market$newspaper),main="new

>

> plot(density(x=log_news),main="newspaper")

> skewness(log_news)

[1] -0.8309352
```



Shapiro-Wilk test

- The third assumption for a linear regression model is that the *residuals/errors* follow a normal distribution.
- So to check whether *residuals* follow a normal distribution we use the Shapiro-Wilk test.
- The Hypothesis for Shapiro-Wilk test are:

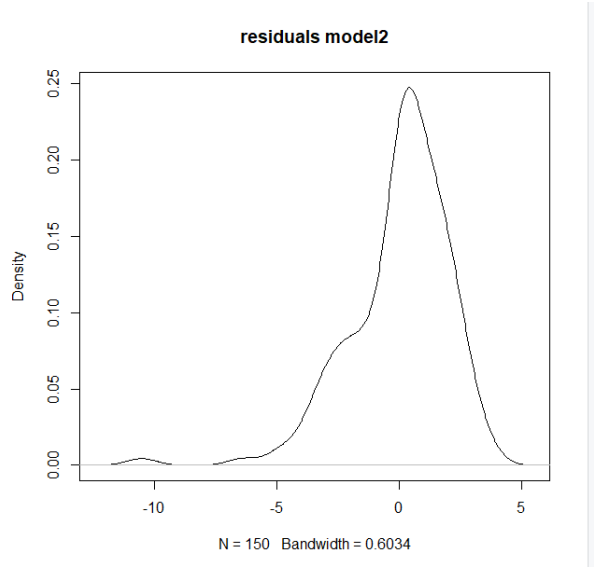
$$H0: Residuals\ follow\ a\ Normal\ Distribution$$
$$vs\ H1: Residuals\ do\ not\ follow\ a\ Normal\ Distribution$$

- If the *p-value* of this test is less than 5%, we reject $H0$ and conclude that residuals do not follow a normal distribution.
- Since we want the *residuals* to follow normal distribution , we want the *p-value* to be greater than 0.05
- This test can be conducted in R by using the ***shapiro.test()*** function from the *stats* package.
- This test can only be conducted for a sample size of less than 5000.

Shapiro-Wilk test

Output:

```
> library(stats)
> shapiro.test(model2$residuals)

        Shapiro-Wilk normality test

data:  model2$residuals
W = 0.91673, p-value = 1.313e-07
dev.off()
plot(density(x=model2$residuals),main="residuals model2")
```

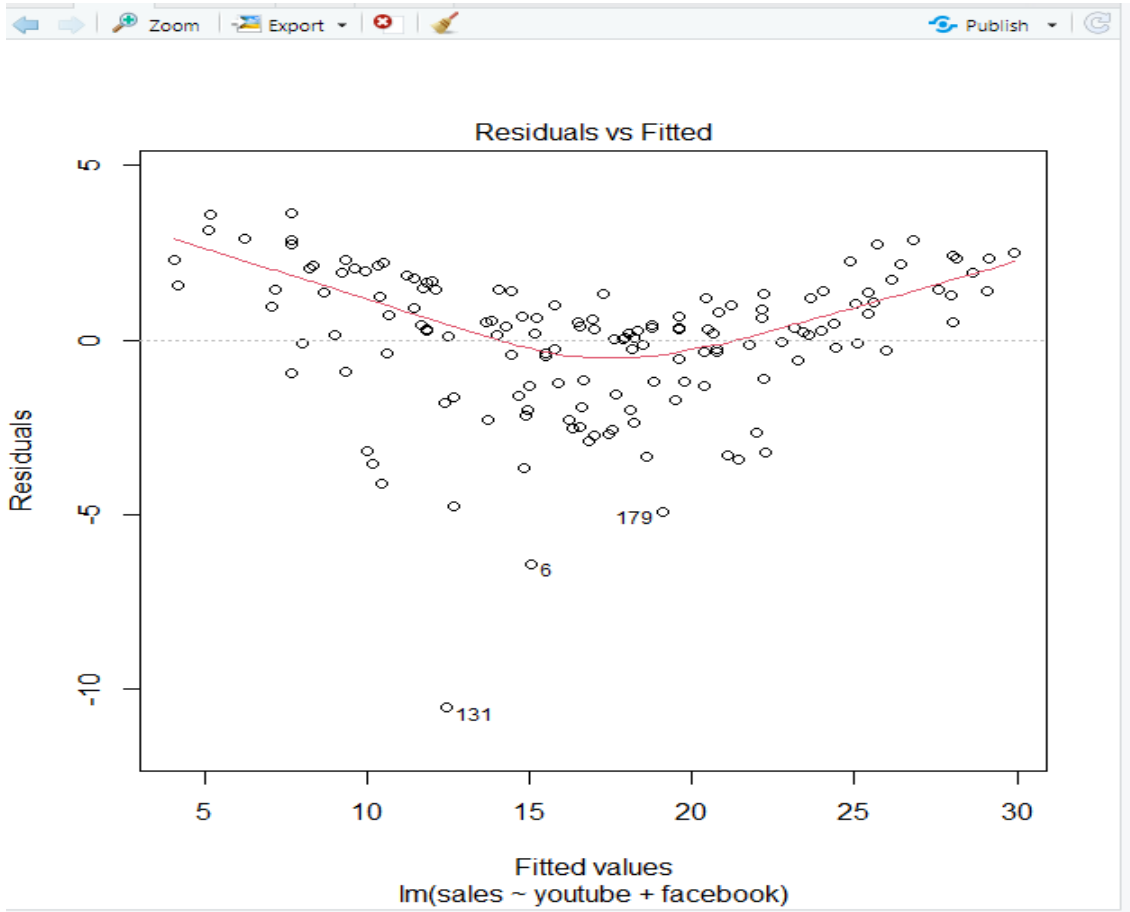


- Since *p-value* is less than 0.05, we reject *H*0 and infer that the *residuals* do not follow a normal distribution.
- In real life data rarely follow a normal distribution and hence to fit a model with normally distributed *residuals* with 5% level of significance.
- For this reason, we use a Q-Q plot of the residuals and ascertain the seriousness of the issue regarding the normality of *residuals*.

Diagonistic Plots **(residuals plot)**

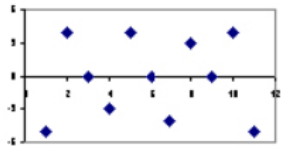
- The first plot has fitted values on the X-axis and residuals on the Y-axis.
- This plot is used to identify any *non-linear* relationship between the independent variables and predicted variable.
- The red dotted line should be horizontal line without any distinct pattern.
- This is a good outcome since it indicates that we don't have any *non-linear* relationships**.**

plot(model2)



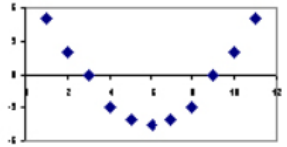
Residual plots

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model.

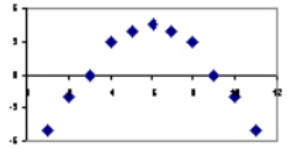


Random pattern

The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.



Non-random: U-shaped curve



Non-random: Inverted U

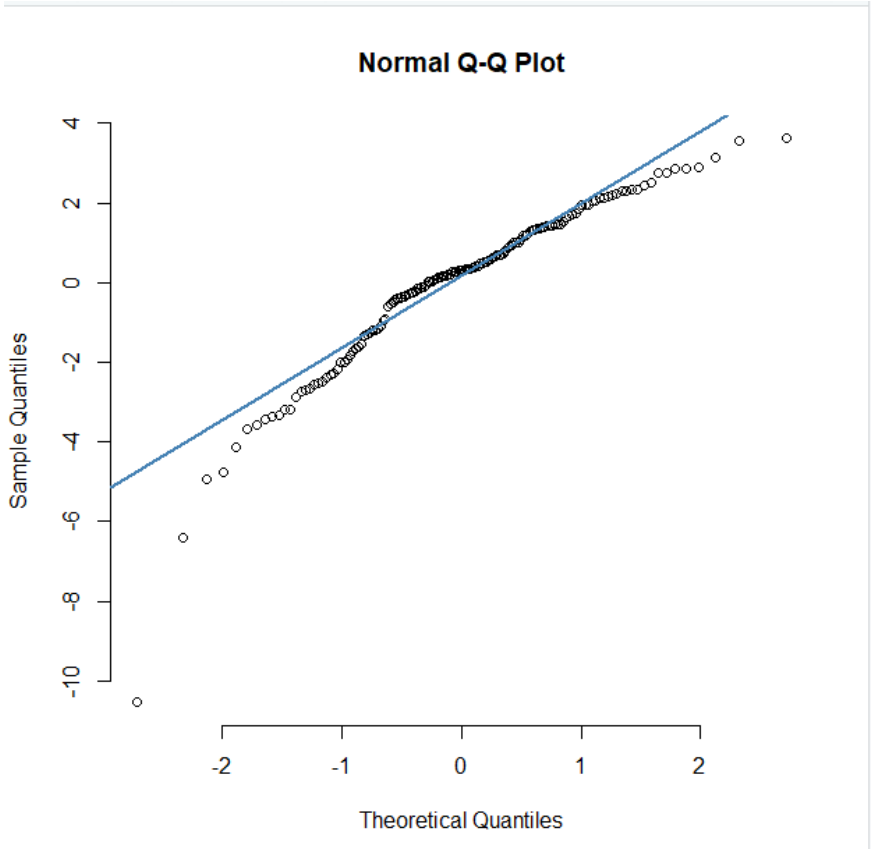
Model Evaluation

- **The Normal Q-Q(quantile-quantile) plot** is designed specially to check for normality of *residuals*.
- If *residuals* are normally distributed the quantiles should form a straight line , i.e. they should be near to the dotted line in the plot.
- In this case, the lower tail(lower left corner) is quite far from the dotted line and the upper tail (top right corner) is also a bit far from the line.
- Here, *residuals* don't seem to follow a Normal Distribution in the tails. The model is indicating three outliers[131,5 and 132]

Lets look at the next plot while keeping in mind about the outliers.

qqnorm(residuals(model2), pch = 1, frame = FALSE)

qqline(residuals(model2), col = "steelblue", lwd = 2)



Multicollinearity

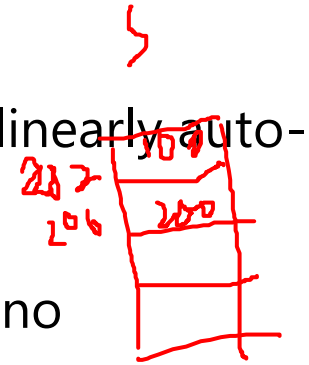
- Linear regression assumes that there is little or no multicollinearity in the data.
- Multicollinearity occurs when the independent variables are too highly correlated with each other.
- Multicollinearity may be tested with three central criteria:
- 1) Correlation matrix – when computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1.
- 3) Variance Inflation Factor (VIF) – the variance inflation factor of the linear regression is defined as VIF = 1/T. With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the variables.
- If multicollinearity is found in the data, centering the data (that is deducting the mean of the variable from each score) might help to solve the problem. However, the simplest way to address the problem is to remove independent variables with high VIF values.

Multicollinearity

```
> library(car)

Loading required package: carData

Warning messages:

1: package ‘car’ was built under R version 4.0.5

2: package ‘carData’ was built under R version 4.0.3

> vif(model2)

 youtube facebook

1.005157 1.005157
```

Autocorrelation

- While a scatterplot allows you to check for autocorrelations, you can test the linear regression model for autocorrelation with the Durbin-Watson test.
- Durbin-Watson's d tests the null hypothesis that the residuals are not linearly auto-correlated.
- While d can assume values between 0 and 4, values around 2 indicate no autocorrelation.
- As a rule of thumb values of $1.5 < d < 2.5$ show that there is no auto-correlation in the data. However, the Durbin-Watson test only analyses linear autocorrelation and only between direct neighbors, which are first order effects.



What is The Durbin Watson Test?

The Durbin Watson Test is a measure of autocorrelation (also called serial correlation) in residuals from regression analysis.

Autocorrelation is the similarity of a time series over successive time intervals.

It can lead to underestimates of the standard error and can cause you to think predictors are significant when they are not.

The Durbin Watson test looks for a specific type of serial correlation, the AR(1) process.

Durbin Watson test

The Hypotheses for the Durbin Watson test are:

H_0 = no first order autocorrelation.

H_1 = first order correlation exists.

(For a first order correlation, the lag is one time unit).

Durbin Watson test

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Where E_t are residuals from an ordinary least squares regression.

The Durbin Watson test reports a test statistic, with a value from 0 to 4, where:

2 is no autocorrelation.

0 to <2 is positive autocorrelation (common in time series data).

>2 to 4 is negative autocorrelation (less common in time series data).

A **rule of thumb** is that test statistic values in the range of 1.5 to 2.5 are relatively normal.

Values outside of this range could be cause for concern.

Autocollinearity test

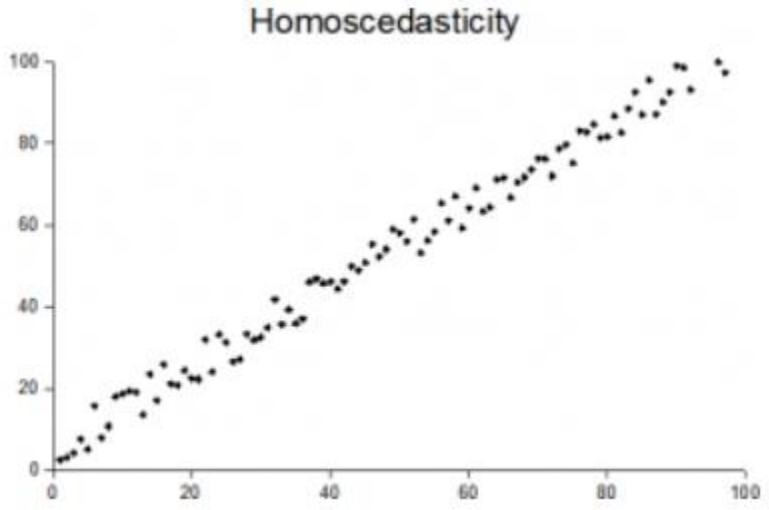
```
> durbinWatsonTest(model2)
 lag Autocorrelation D-W Statistic p-value
   1     -0.05325835       2.104747   0.554
 Alternative hypothesis: rho != 0
```

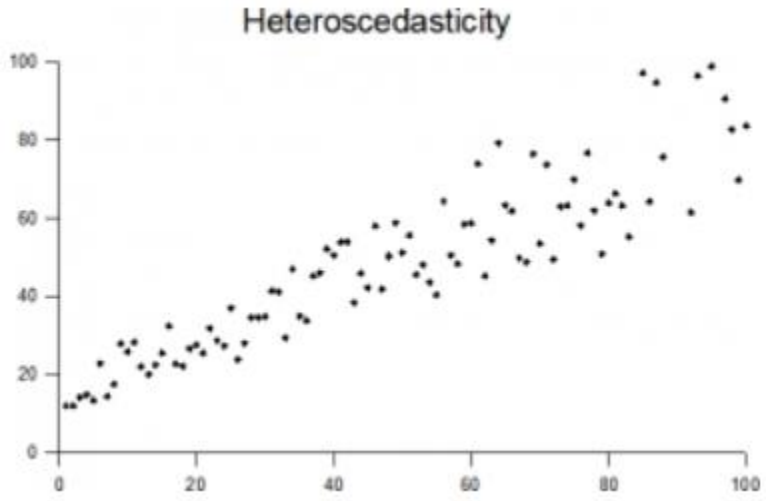
values in the range of 1.5 to 2.5 are relatively normal.

homoscedasticity

Homoscedasticity described as a condition where the standard deviations are equal for all points.

Simply put, **homoscedasticity** means “having the same scatter.” For it to exist in a set of data, the points must be about the same distance from the line, as shown in the picture above. The opposite is *hetero*scedasticity (“different scatter”), where points are at widely





Breusch Pagan test

- Heteroskedasticity means changing/non-constant variance.
- In linear regression, error terms should be homoskedastic i.e. they should have a common constant variance.
- We conduct a Breusch-Pagan test to check for heteroskedasticity among the error terms.
- The Hypothesis of Breusch-Pagan test is:

$$H0: Data\ is\ homoskedastic,\ ie\ error\ variances\ are\ equal$$
$$H1: Data\ is\ heteroskedastic$$

- We reject $H0$ only if *p-value* is less than 5%.
- Here, we want the errors to be homoscedastic and so we want to have a *p-value* to be greater than 5%
- In R, the Breusch-Pagan test is conducted using the ***bptest()*** function available in the *lmtest* package.

Breusch Pagan test

Input:

```
install.packages("lmtest")
library(lmtest)
bptest(step.model)
```

Output:

```
        studentized Breusch-Pagan test

data:  step.model
BP = 1.707, df = 2, p-value = 0.4259
```

Inference:

Since the *p-value>0.05* , we don't have enough evidence to reject $H0$

We conclude that the error terms are homoscedastic in nature.

- The term inverse can be used with different meanings. The meanings are: reciprocal. In this case the inverse of log(x) is 1/log(x) inverse function.
- In this case it refers to solving the equation log(y) = x for y
- in which case the inverse transformation is exp(x) assuming the log is base e.
- (In general, the solution is b^x if the log is of base b.
- For example, if log10(y) = x then the inverse transformation is 10^x.)

Shapiro-Wilk test

- The third assumption for a linear regression model is that the *residuals/errors* follow a normal distribution.
- So to check whether *residuals* follow a normal distribution we use the Shapiro-Wilk test.
- The Hypothesis for Shapiro-Wilk test are:

$$H0: Residuals\ follow\ a\ Normal\ Distribution$$

$$vs\ H1: Residuals\ do\ not\ follow\ a\ Normal\ Distribution$$

- If the *p-value* of this test is less than 5%, we reject $H0$ and conclude that residuals do not follow a normal distribution.
- Since we want the *residuals* to follow normal distribution , we want the *p-value* to be greater than 0.05
- This test can be conducted in R by using the ***shapiro.test()*** function from the *stats* package.
- This test can only be conducted for a sample size of less than 5000.

Shapiro-Wilk test

Input:

```
#test for normality of residuals
#shapiro-wilk test
library("stats")
shapiro.test(step.model$residuals)
```

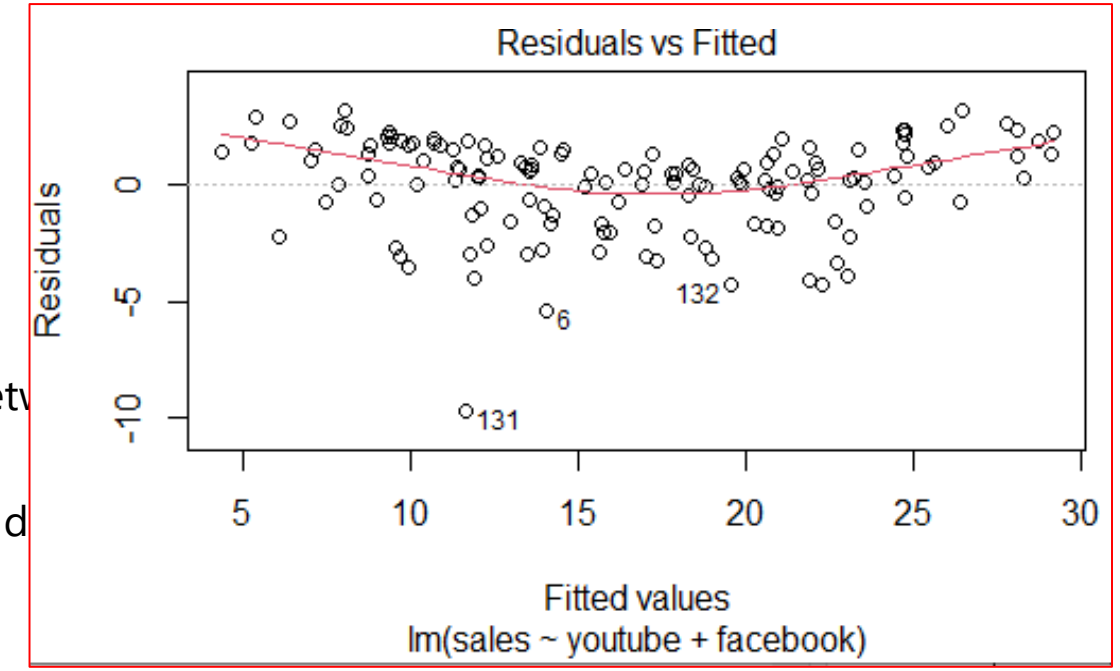
Output:

```
> shapiro.test(step.model$residuals)

        Shapiro-Wilk normality test

data:  step.model$residuals
W = 0.91598, p-value = 2.628e-07
```

- Since *p-value* is less than 0.05, we reject $H0$ and infer that the *residuals* do not follow a normal distribution.
- In real life data rarely follow a normal distribution and hence to fit a model with normally distributed *residuals* with 5% level of significance.
- For this reason, we use a Q-Q plot of the residuals and ascertain the seriousness of the issue regarding the normality of *residuals*.

Diagonistic Plots

```
> plot(step.model)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```

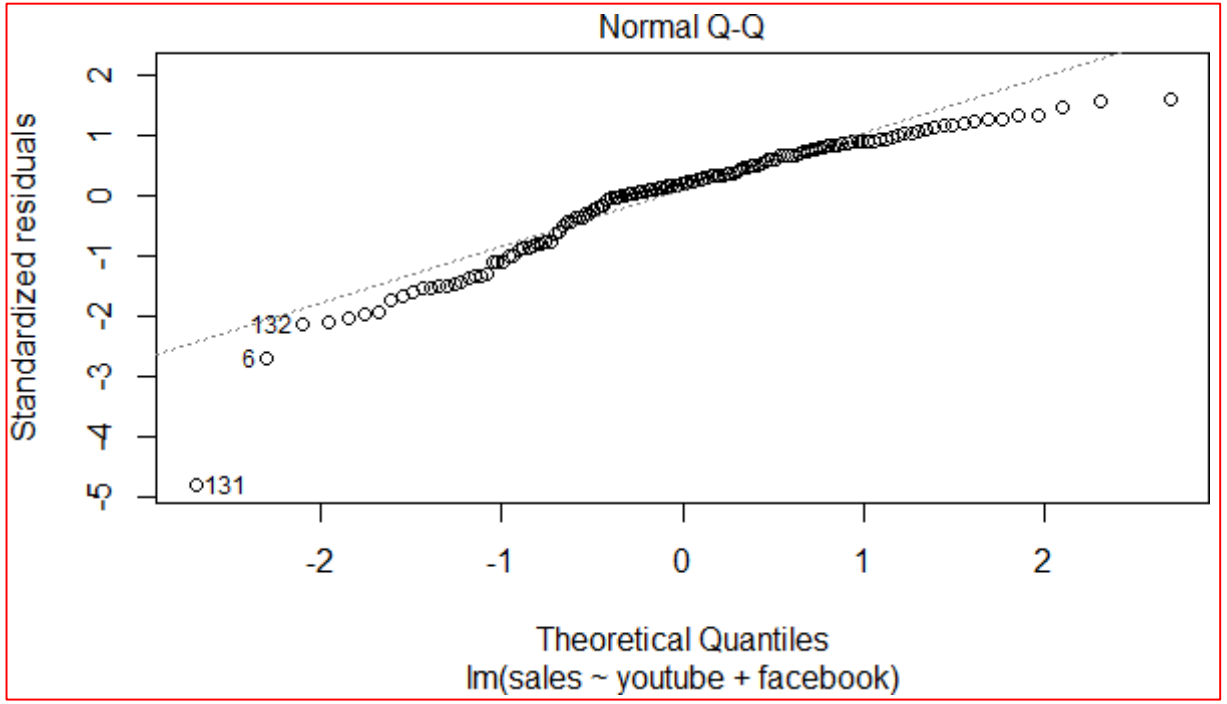


- The first plot has fitted values on the X-axis and residuals on the Y-axis.
- This plot is used to identify any *non-linear* relationship betw variable.
- The red dotted line should be horizontal line without any d
- This is a good outcome since it indicates that we don't have any *non-linear* relationships.

Model Evaluation

- **The Normal Q-Q(quantile-quantile) plot** is designed specially to check for normality of *residuals*.
- If *residuals* are normally distributed the quantiles should form a straight line , i.e. they should be near to the dotted line in the plot.
- In this case, the lower tail(lower left corner) is quite far from the dotted line and the upper tail (top right corner) is also a bit far from the line.
- Here, *residuals* don't seem to follow a Normal Distribution in the tails. The model is indicating three outliers[131,5 and 132]

Lets look at the next plot while keeping in mind about the outliers.

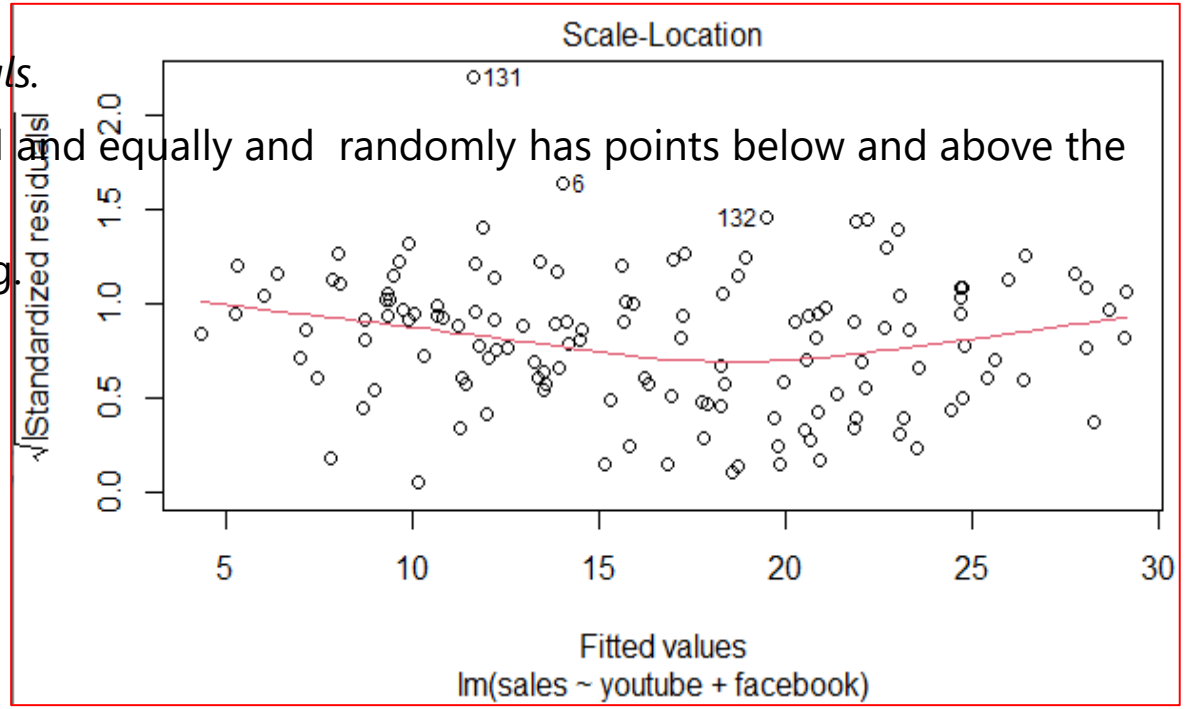


Model Evaluation

➢The third diagnostic plot has the fitted values on the X-axis and standardized residuals on the Y-axis. This plot is called a *Scale-Location* or *Spread-Location* plot.

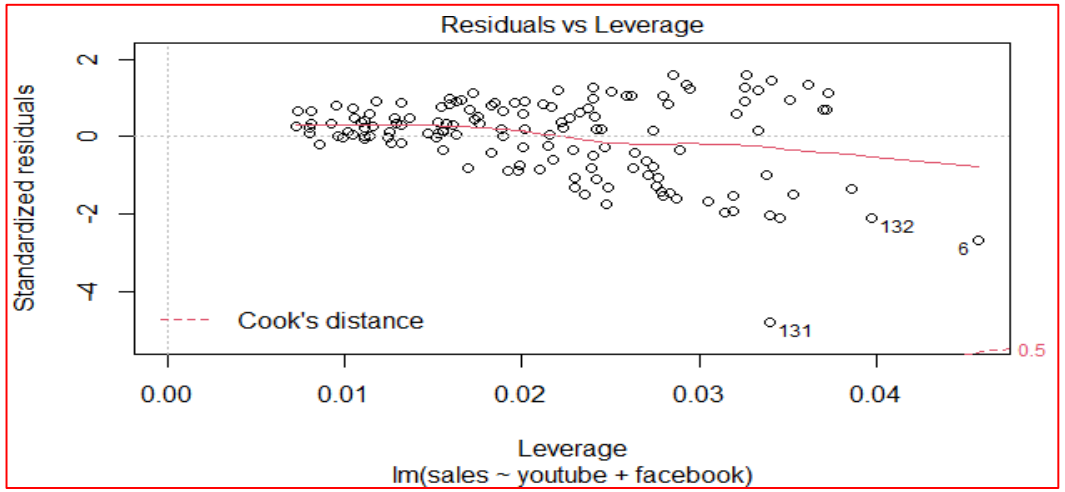
- This plot helps in checking the homoskedasticity of the *residuals*.
- A good outcome from this plot is that the red line is horizontal and equally and randomly has points below and above the line.
- The assumption of homoskedasticity fails if the red line is rising.
- Here the red line is horizontal and has almost

  equal points on both sides.
- *Residuals* possess homoskedasticity.



Model Evaluation

- The fourth plot is of **standardized residuals versus leverage.**
- Leverage implies the impact of data points on the model.
- Higher the leverage, higher the impact.
- This plot is used to check if the outliers mentioned in the second plot significantly influence the model.
- Here, the pattern of the red line is not relevant.
- The relevant part is the ***Cook's Distance***. If the outliers lie beyond this distance we conclude that they significantly impact our model and should be removed to improve the model.
- Here no outliers lie beyond that distance[marked at lower right corner] and so we conclude that outliers don't significantly impact our regression model.



Thank you

1.2 Icons to use

	To highlight something important
	To ask a question
	When giving a reference to extra/additional reading
	Question to be solved (in class)
	Important definition
	To quote someone