IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Estimation, Prediction of Regression Model Residual Analysis: Validating Model Assumptions - II

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- Understanding different types of residual analysis
- Plotting residual plots using python

Residual Analysis: Validating Model Assumptions

- **Residual analysis** *is the primary tool for determining whether the assumed regression model is appropriate*

RESIDUAL FOR OBSERVATION i

$$y_i - \hat{y}_i$$

where

y_i is the observed value of the dependent variable
$\hat{y}_i$ is the estimated value of the dependent variable

Assumptions about the error term . ξ

$$y = \beta_0 + \beta_1 x + \epsilon$$

1. $E(\epsilon) = 0$.
2. The variance of ϵ, denoted by σ^2, is the same for all values of x.
3. The values of ϵ are independent.
4. The error term ϵ has a normal distribution.

Importance of the Assumptions

- These assumptions provide the theoretical basis for the *t* test and the *F* test used to determine whether the relationship between *x* and *y* is significant, and for the confidence and prediction interval estimates
- If the assumptions about the error term ξ appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

Residuals for Ice cream parlours

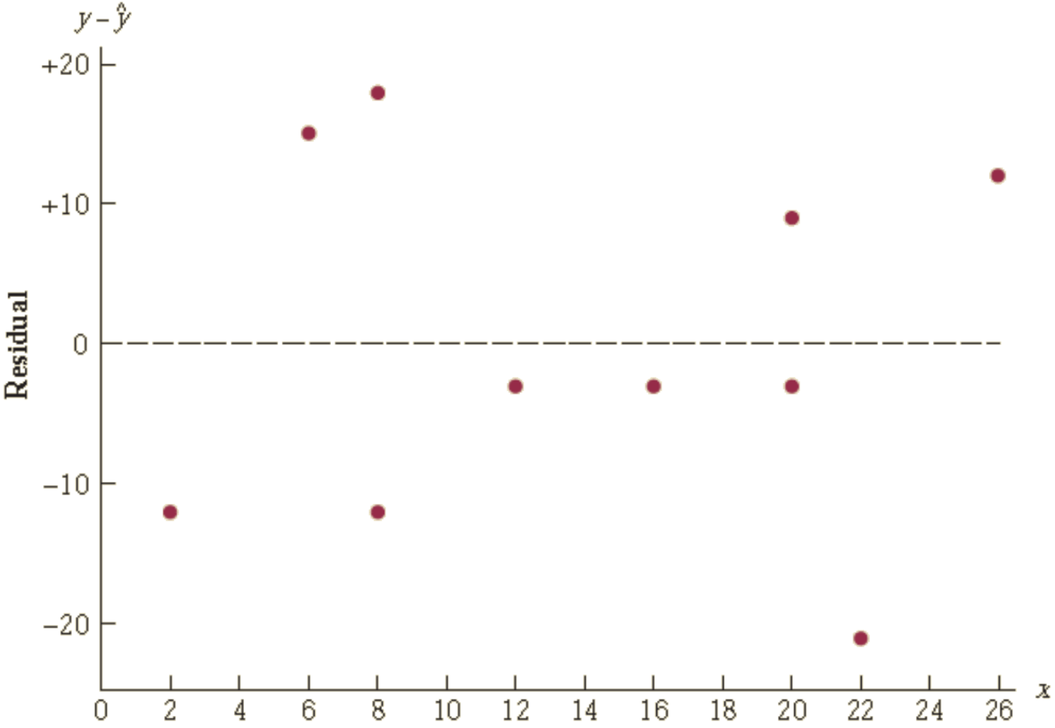
Student Population x_i	Sales y_i	Estimated Sales $\hat{y}_i = 60 + 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

Source: Statistics for Business & Economics, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, Cengage Learning,2013

Residual analysis is based on an examination of graphical plots

- A plot of the residuals against values of the independent variable x
- A plot of residuals against the predicted values of the dependent variable $\hat{y}$
- A standardized residual plot
- A normal probability plot

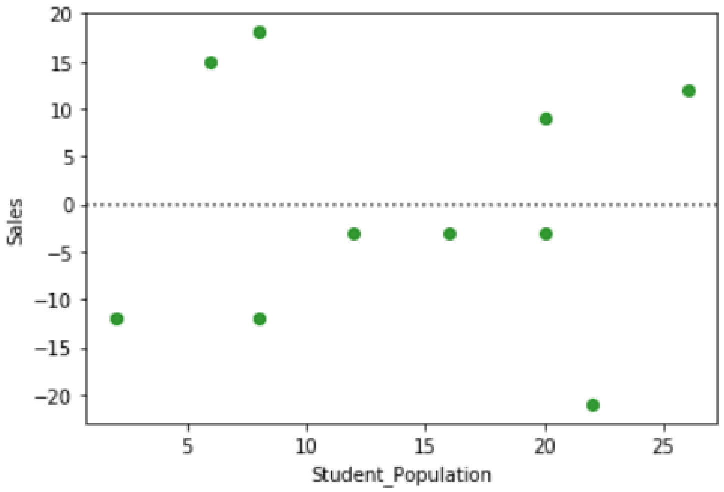
Residual Plot Against *x*



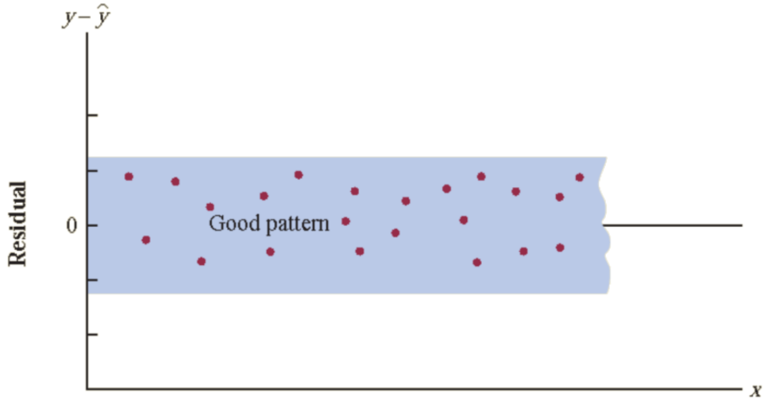
Residual Plot Against *x*

```
In [18]: import seaborn as sns
         sns.residplot(df1['Student_Population'],df1['Sales'], color="g")

Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x24e594e9f60>
```

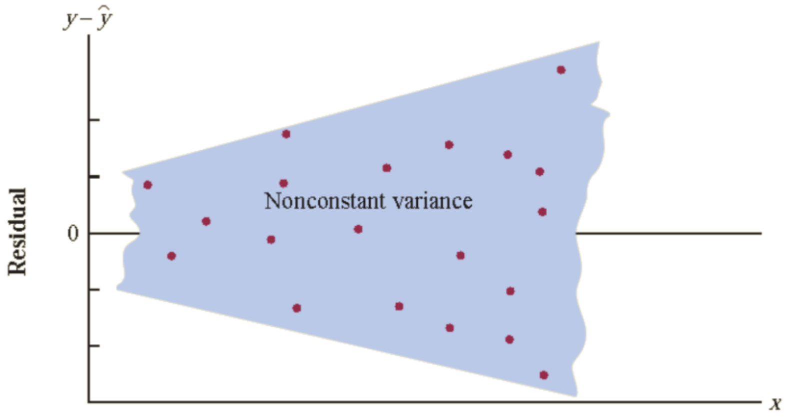


Assumption: the variance is the same for all values of *x*



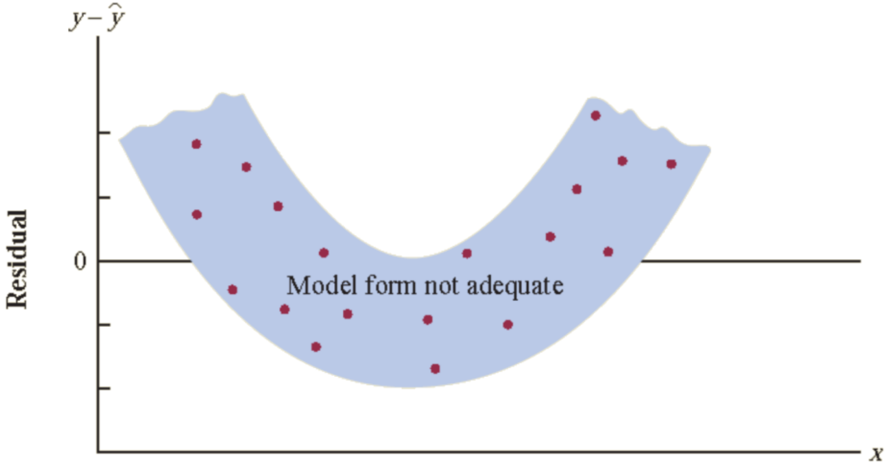
- The residual plot should give an overall impression of a horizontal band of points

Violation of Assumption:
The variance of ‘e’ is not the same for all values of *x*



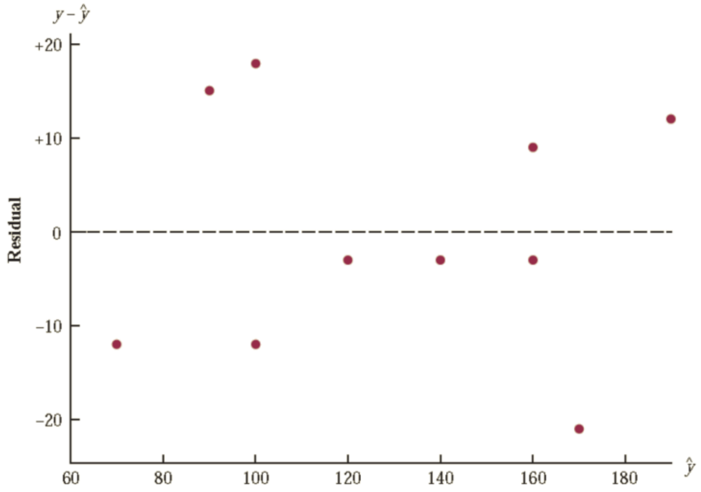
- Assumption of a constant variance of ‘e’ is violated
- If variability about the regression line is greater for larger values of *x*

Assumed regression model is not an adequate representation



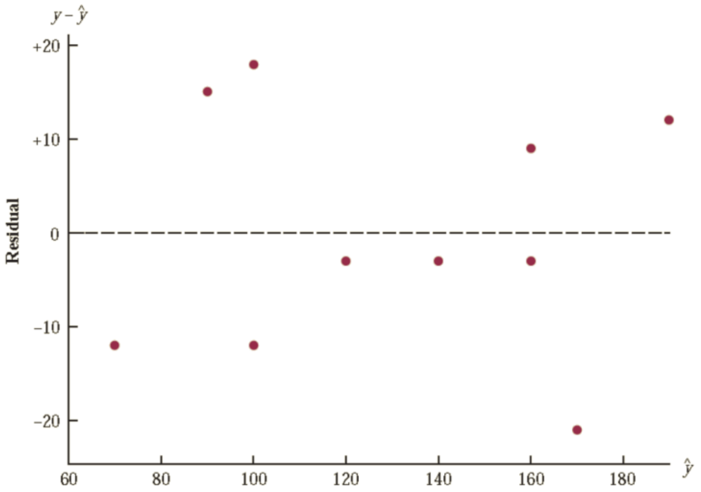
A curvilinear regression model or multiple regression model should be considered.

Residual Plot Against $\hat{y}$



- The pattern of this residual plot is the same as the pattern of the residual plot against the independent variable *x*.
- It is not a pattern that would lead us to question the model assumptions.

Residual Plot Against $\hat{y}$



- For simple linear regression, both the residual plot against *x* and the residual plot against provide the same pattern
- For multiple regression analysis, the residual plot against $\hat{y}$ is more widely used because of the presence of more than one independent variable.

Standardized Residuals

- Many of the residual plots provided by computer software packages use a standardized version of the residuals.
- A random variable is standardized by subtracting its mean and dividing the result by its standard deviation.
- With the least squares method, the mean of the residuals is zero.
- Thus, simply dividing each residual by its standard deviation provides the **standardized residual**

Python Code

In [14]:
```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import matplotlib.pyplot as plt
```

In [9]:
```
df1 = pd.read_excel('Icecream.xlsx')
df1
```

Out[9]:

	Student_Population	Sales
0	2	58
1	6	105
2	8	88
3	8	118
4	12	117
5	16	137
6	20	157
7	20	169
8	22	149

In [11]:
```
Reg1 = ols(formula = "Sales ~ Student_Population", data = df1)
Fit1 = Reg1.fit()
print(Fit1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.891
Method:                 Least Squares   F-statistic:                     74.25
Date:                Thu, 05 Sep 2019   Prob (F-statistic):           2.55e-05
Time:                        11:16:42   Log-Likelihood:                -39.342
No. Observations:                  10   AIC:                             82.68
Df Residuals:                       8   BIC:                             83.29
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         60.0000      9.226      6.503      0.000      38.725      81.275
x1             5.0000      0.580      8.617      0.000       3.662       6.338
==============================================================================
Omnibus:                        0.928   Durbin-Watson:                   3.224
Prob(Omnibus):                  0.629   Jarque-Bera (JB):                0.616
Skew:                          -0.060   Prob(JB):                        0.735
Kurtosis:                       1.790   Cond. No.                         33.6
==============================================================================
```

Python Code

```
In [12]: print(anova_lm(Fit1))

                    df   sum_sq   mean_sq          F    PR(>F)
Student_Population  1.0  14200.0  14200.00  74.248366  0.000025
Residual            8.0   1530.0    191.25        NaN       NaN
```

Standardized Residuals

STANDARD DEVIATION OF THE ith RESIDUAL

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i}$$

where

$$s_{y_i - \hat{y}_i} = \text{the standard deviation of residual } i$$

$$s = \text{the standard error of the estimate}$$

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

Computation of standardized residuals for Icecream parlors

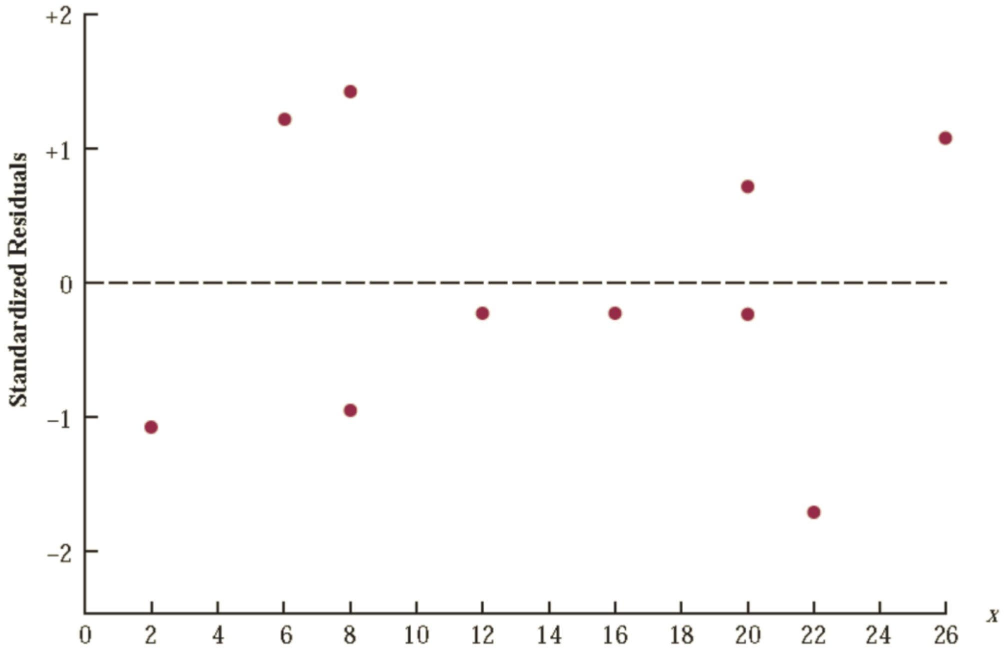
Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	−12	144	.2535	.3535	11.1193	−12	−1.0792
2	6	−8	64	.1127	.2127	12.2709	15	1.2224
3	8	−6	36	.0634	.1634	12.6493	−12	−.9487
4	8	−6	36	.0634	.1634	12.6493	18	1.4230
5	12	−2	4	.0070	.1070	13.0682	−3	−.2296
6	16	2	4	.0070	.1070	13.0682	−3	−.2296
7	20	6	36	.0634	.1634	12.6493	−3	−.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	−21	−1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
		Total	568					

Computation of standardized residuals for Icecream parlors

STANDARDIZED RESIDUAL FOR OBSERVATION i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$

Plot of The Standardized Residuals Against The Independent Variable x



Plot of The Standardized Residuals Against The Independent Variable *x*

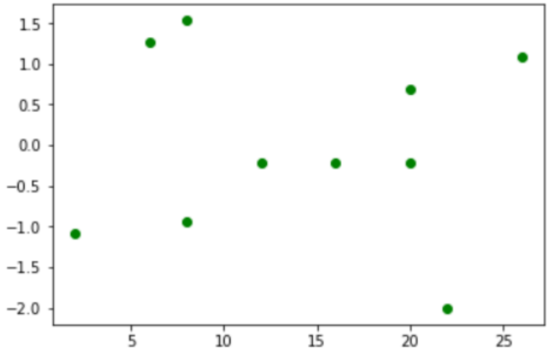
In [21]:
```
influence =Fit1.get_influence()
resid_student = influence.resid_studentized_external
```

In [22]:
```
resid_student
```

Out[22]:
```
array([-1.09212653,  1.26798654, -0.94196706,  1.54023214, -0.21544891,
       -0.21544891, -0.22263461,  0.68766487, -2.01063738,  1.09212653])
```

In [24]:
```
plt.figure()
plt.scatter(df1['Student_Population'],resid_student, color = "green")
```

Out[24]:
```
<matplotlib.collections.PathCollection at 0x24e5a382b38>
```



Studentized residual

- The standardized residual plot can provide insight about the assumption that the error ‘e’ term has a normal distribution.
- If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.

Studentized residual

- Thus, when looking at a standardized residual plot, we should expect to see approximately 95% of the standardized residuals between -2 and 2.
- We see in Figure that for the Armand's example all standardized residuals are between -2 and 2.
- Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that 'e' has a normal distribution.

Normal Probability Plot

- Another approach for determining the validity of the assumption that the error term has a normal distribution is the **normal probability plot**.
- To show how a normal probability plot is developed, we introduce the concept of *normal scores.*

Normal Probability Plot

- Suppose 10 values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of 10 ordered from smallest to largest.
- For now, let us consider only the smallest value in each sample.
- The random variable representing the smallest value obtained in repeated sampling is called the first-order statistic.

Normal Probability Plot

NORMAL SCORES
FOR $n = 10$

Order Statistic	Normal Score
1	−1.55
2	−1.00
3	−.65
4	−.37
5	−.12
6	.12
7	.37
8	.65
9	1.00
10	1.55

Normal Probability Plot

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	−12	144	.2535	.3535	11.1193	−12	−1.0792
2	6	−8	64	.1127	.2127	12.2709	15	1.2224
3	8	−6	36	.0634	.1634	12.6493	−12	−.9487
4	8	−6	36	.0634	.1634	12.6493	18	1.4230
5	12	−2	4	.0070	.1070	13.0682	−3	−.2296
6	16	2	4	.0070	.1070	13.0682	−3	−.2296
7	20	6	36	.0634	.1634	12.6493	−3	−.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	−21	−1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
		Total	568					

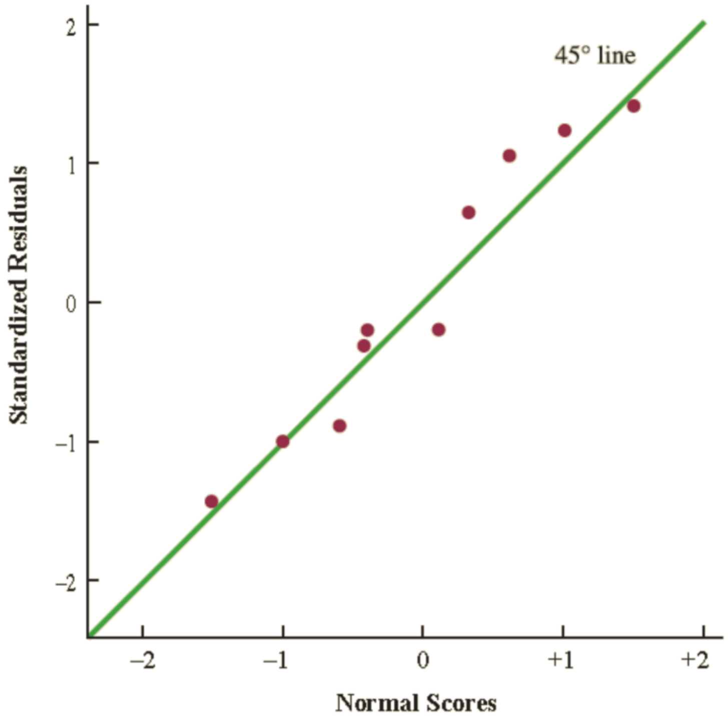
Normal scores and ordered standardized residuals for Armand's pizza parlors

Normal Scores	Ordered Standardized Residuals
−1.55	−1.7114
−1.00	−1.0792
−.65	−.9487
−.37	−.2372
−.12	−.2296
.12	−.2296
.37	.7115
.65	1.0792
1.00	1.2224
1.55	1.4230

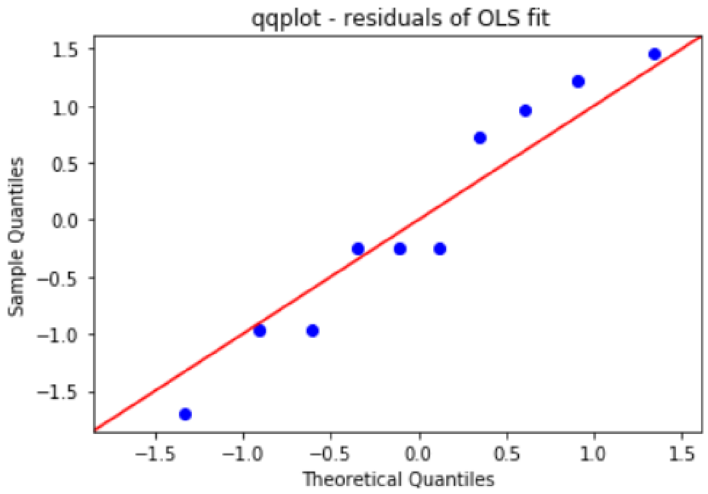
Normal Probability Plot

- If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest normal score, and so on.
- If we were to develop a plot with the normal scores on the horizontal axis and the corresponding standardized residuals on the vertical axis, the plotted points should cluster closely around a 45-degree line passing through the origin if the standardized residuals are approximately normally distributed.
- Such a plot is referred to as a *normal probability plot.*

Normal probability plot for Ice Cream parlors



In [29]:
```
from scipy import stats
import statsmodels.api as sm
res = Fit1.resid # residuals
probplot = sm.ProbPlot(res,stats.norm, fit=True)
fig = probplot.qqplot(line='45')
h = plt.title(' qqplot - residuals of OLS fit')
plt.show()
```



Thank You