625.661 - Homework Four

Eric Niblock

March 14, 2022

1. **Randomly select 20 rows of Table B.5 of Textbook. Then perform a multiple regression fit to the data you generated. The multiple regression model contains the response variable y (CO2) and regressors x_1 (space time in min) and x_6 (solvent total) and intercept.**

   (a) **Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?**

   The normal probability plot is provided in the attached PDF. There does not appear to be an issue with the normality assumption, as the plot appears to be approximately linear.

   (b) **Construct and interpret a plot of the residuals versus the predicted response.**

   A plot of the residuals vs the response, y, is constructed in the attached PDF. There does appear to be an issue with the assumption of constant variance. The plot suggests that the variance grows in proportion to the response, as evident by the cone shape.

   (c) **Compute the studentized residuals and the R-student residuals for this model. What information is conveyed by these scaled residuals?**

   The studentized residuals are given below, and are formally calculated in the attached PDF,

```
array([-0.07814444,  1.20726661,  0.09619435,  2.09537083, -0.8214521 ,
       -0.17027085, -1.86087022,  0.29133745, -0.9105145 ,  0.33971069,
       -0.87592948,  0.66878824, -0.68064628, -0.78315718, -0.00591975,
        0.42784005,  1.5883014 , -0.16263111,  1.31425049, -1.55955612])
```

Additionally, the R-student residuals are given below, and are formally calculated in the attached PDF,

```
array([-0.07595563,  1.22384659,  0.09350815,  2.3418848 , -0.81370572,
       -0.16560695, -2.01233709,  0.28379901, -0.90597037,  0.33120287,
       -0.86999416,  0.65817419, -0.67014957, -0.77439984, -0.00575296,
        0.41791617,  1.66459965, -0.15816527,  1.34329744, -1.62971649])
```

Both the studentized residuals and the R-student residuals are standardized residuals in that they are scaled by standard deviations specific to each residual, e_i, rather than the overall MS_{Res} (without adjustment). The only difference between these two types of residuals lies in the value of MS_{Res} used in the calculation of the standard deviation. For studentized residuals, the value of MS_{Res} is calculated as normal, while for R-student residuals the value of MS_{Res} is calculated on a model that doesn't have the point associated with the residual in question.

Any studentized or R-student residuals that have a magnitude greater than 3 could be classified as outliers, however in this example, we have no such residuals. Furthermore, it is generally the case that studentized and R-student residuals are approximately equal. Any large discrepancies between the two would indicate that the observation is particularly influential on model behavior. Here, we have no such discrepancies. These residuals help to confirm the normality assumption.

(d) Compute all other residuals (e.g., PRESS) to examine whether there are some observations that may not fit the model or potential outliers.

The PRESS residuals are given below, and are formally calculated in the attached PDF,

```
array([ -0.67868144,  10.55601826,   0.84255854,  18.29758329,
        -7.13853966,  -1.50046544, -16.26516822,   2.55234556,
        -8.56706985,   3.11426965,  -7.61088844,   5.87050615,
```

```
-5.92525147,  -6.91698176,  -0.05625853,   3.74062033,
14.79590205,  -1.58592958,  14.1173234 , -15.21928743])
```

A case could be made to conclude that the fourth observation, associated with the PRESS residual of 18.298, is an outlier. This residual has a corresponding h_{ii} of 0.97, which then inflates the PRESS residual. In other words, the observation is highly influential on the model.

2. Randomly select 15 rows of Table B.4 (Property Valuation Data) of Textbook.

(a) Perform a thorough regression analysis of y on x_4, x_7, and x_9, including residual plots.

The attached PDF contains the multiple regression model, as well as the residual plots. It appears that the normality assumption holds, though the residuals are linearly related to the response, y.

(b) Can an appropriate test for lack of fit be constructed? Why or why not?

An appropriate lack of fit test cannot be constructed. A lack of fit test requires us to have multiple observation of y given repeated values of $\mathbf{x}$, where $\mathbf{x}$ in this case is given by $\mathbf{x} = (x_4, x_7, x_9)$. There are no instances where $\mathbf{x_i} = \mathbf{x_j}$ for $i, j = 1, ..., n$.

3. Randomly select 7 rows of the data in Problem 5.5 of Textbook. Then complete Problem 5.5: A glass bottle manufacturing company has recorded data on the average number of defects per 10,000 bottles due to stones (small pieces of rock embedded in the bottle wall) and the number of weeks since the last furnace overhaul.

(a) Fit a straight - line regression model to the data and perform the standard tests for model adequacy.

The attached PDF contains a linear regression as well as residual plots for the randomly selected data. It appears that the residuals are approximately normal, though they do vary somewhat from the linear trend displayed. Additionally, the residuals as a function of the response do not appear to be evenly distributed around the zero line. Therefore, the model may benefit from a transformation.

(b) Suggest an appropriate transformation to eliminate the problems encountered in part (a). Fit the transformed model and check for adequacy.

As seen in the attached PDF, we employed the transformation $y' = ln(y)$ and refit the model. Then, recreating the residual plots reveals that the residuals are approximately normal, and many of the residuals are pulled closer to the zero line when plotted against the response. Furthermore, it is possible that two outliers exist, as seen on the residual vs. response graph.

In [139]:
```
import statsmodels.api as sm
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import scipy.stats as stats
```

Problem 1 - Data Selection

In [ ]:
```
n = 20

df = pd.read_excel(r'C:\Users\maste\Downloads\linear_regression_5e_data_sets\line
                   '\Appendices\data-table-B5.XLS')
sample = df.sample(n)
X = np.array(sample[['x1','x6']])
y = np.array(sample[['y']])
```

In [71]:
```
sample[['y','x1','x6']]
```

Out[71]:

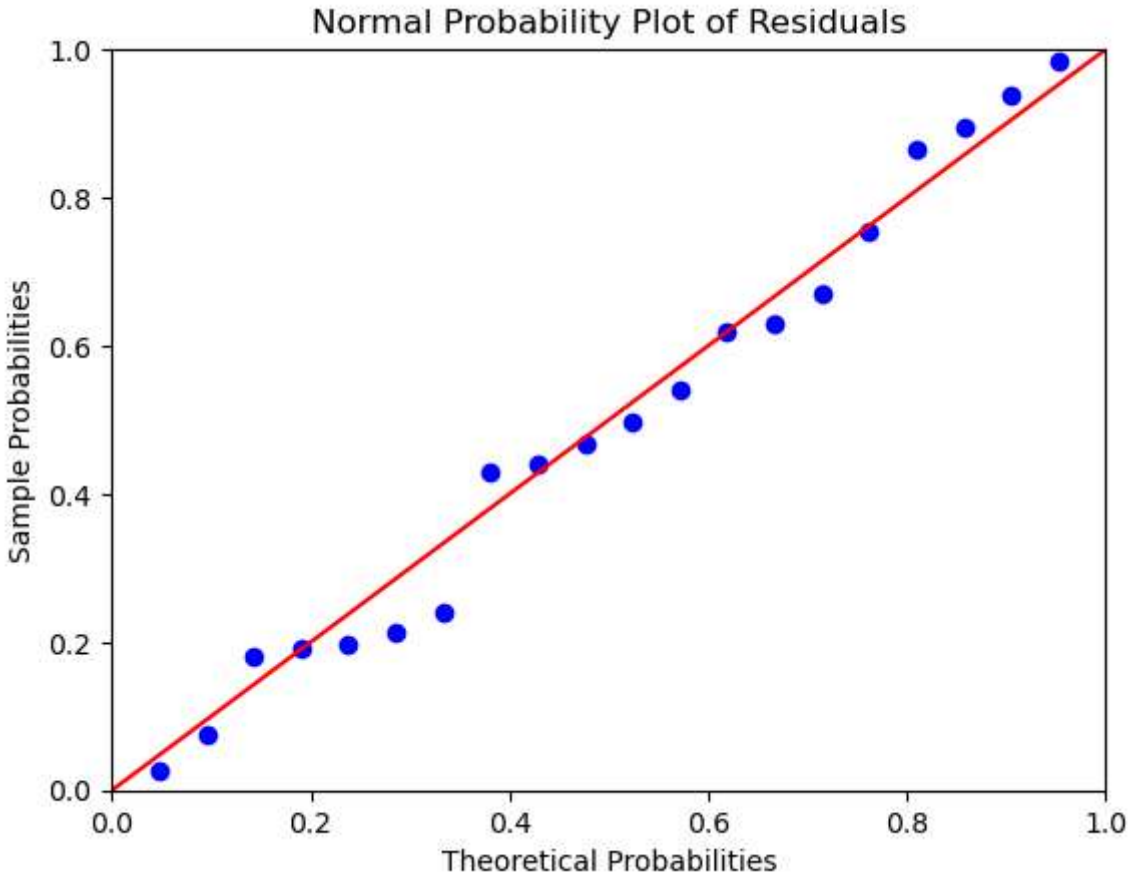
	y	x1	x6
7	35.9	5.8282	6
14	40.5	7.7841	6
15	43.9	9.0384	7
12	36.9	8.2464	8
17	37.9	7.5422	6
20	38.9	8.3607	8
5	30.9	5.8980	7
0	29.5	5.0208	7
13	41.9	6.6969	7
21	36.9	8.1400	7
8	31.5	5.3003	6
6	28.9	5.6039	6
23	25.9	4.9176	7
11	30.0	5.0500	5
19	37.9	6.0831	6

In [19]:
```
X = sm.add_constant(X)
mod = sm.OLS(y, X)
results = mod.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.746
Model:                            OLS   Adj. R-squared:                  0.716
Method:                 Least Squares   F-statistic:                     24.96
Date:                Mon, 14 Mar 2022   Prob (F-statistic):           8.75e-06
Time:                        12:21:25   Log-Likelihood:                -69.130
No. Observations:                  20   AIC:                             144.3
Df Residuals:                      17   BIC:                             147.2
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.2079      8.610      1.186      0.252      -7.959      28.374
x1            -0.1273      0.218     -0.585      0.566      -0.586       0.332
x2             0.0173      0.005      3.677      0.002       0.007       0.027
==============================================================================
Omnibus:                        0.265   Durbin-Watson:                   2.337
Prob(Omnibus):                  0.876   Jarque-Bera (JB):                0.299
Skew:                           0.228   Prob(JB):                        0.861
Kurtosis:                       2.610   Cond. No.                     5.29e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 5.29e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

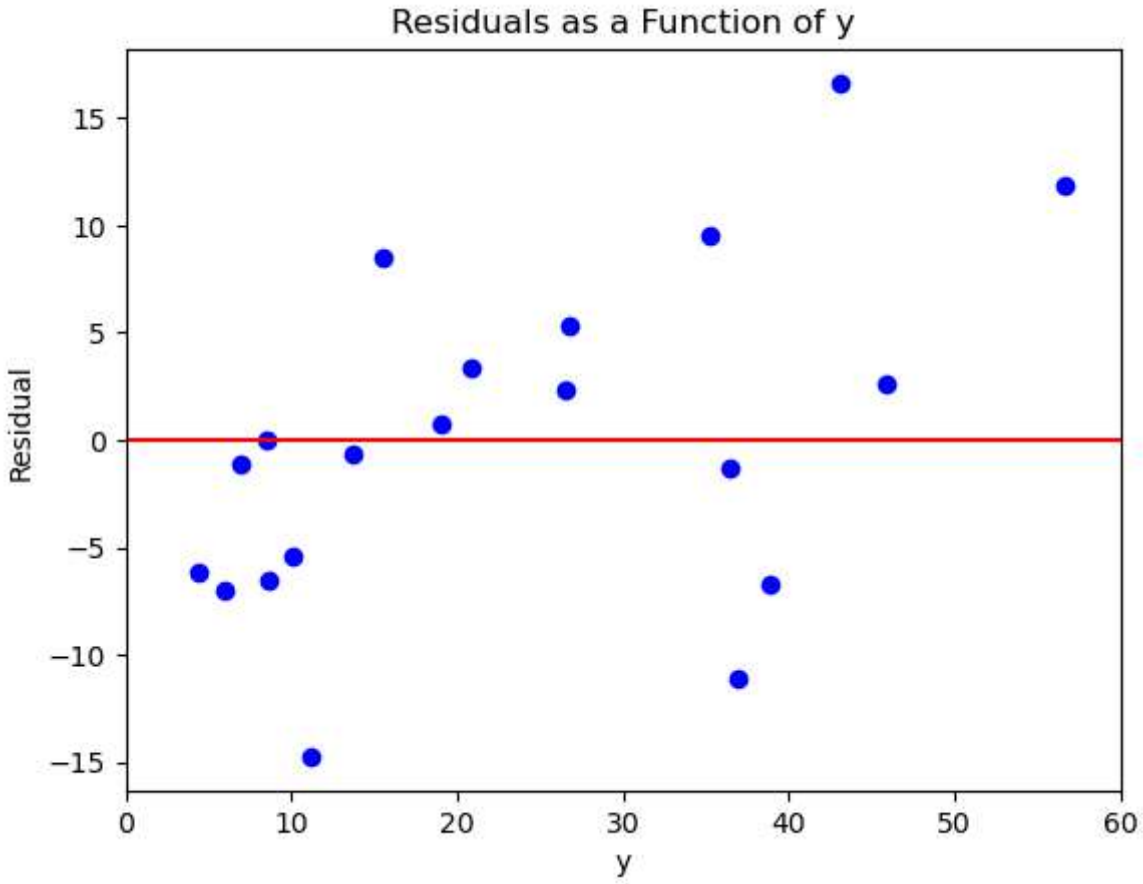
Problem 1(a)

In [62]:
```
res = results.resid
pplot = sm.ProbPlot(res, stats.t, fit=True)
fig = pplot.ppplot(line="45")
h = plt.title("Normal Probability Plot of Residuals")
plt.show()
```



Problem 1(b)

In [63]:
```
plt.scatter(y,res, c='b')
plt.plot([0,60],[0,0],c='r')
plt.xlim(0,60)
plt.xlabel('y')
plt.ylabel('Residual')
h = plt.title("Residuals as a Function of y")
plt.show()
```



Problem 1(c)

In [47]:
```
H = X@np.linalg.inv(X.T@X)@X.T
lev = np.diagonal(H)
MS_RES = results.mse_resid
std_ei = (MS_RES*(1-lev))**0.5
```

In [54]:
```
student_res = res/std_ei
student_res
```

Out[54]:
```
array([-0.07814444,  1.20726661,  0.09619435,  2.09537083, -0.8214521 ,
       -0.17027085, -1.86087022,  0.29133745, -0.9105145 ,  0.33971069,
       -0.87592948,  0.66878824, -0.68064628, -0.78315718, -0.00591975,
        0.42784005,  1.5883014 , -0.16263111,  1.31425049, -1.55955612])
```

In [52]:
```
s2i = ((n-2)*MS_RES - res**2/(1-lev))/(n-3)
```

In [55]:
```
r_student_res = res/((s2i*(1-lev))**0.5)
r_student_res
```

Out[55]:
```
array([-0.07595563,  1.22384659,  0.09350815,  2.3418848 , -0.81370572,
       -0.16560695, -2.01233709,  0.28379901, -0.90597037,  0.33120287,
       -0.86999416,  0.65817419, -0.67014957, -0.77439984, -0.00575296,
        0.41791617,  1.66459965, -0.15816527,  1.34329744, -1.62971649])
```

Problem 1(d)

In [57]:
```
PRESS_res = res/(1-lev)
PRESS_res
```

Out[57]:
```
array([ -0.67868144,  10.55601826,   0.84255854,  18.29758329,
        -7.13853966,  -1.50046544, -16.26516822,   2.55234556,
        -8.56706985,   3.11426965,  -7.61088844,   5.87050615,
        -5.92525147,  -6.91698176,  -0.05625853,   3.74062033,
        14.79590205,  -1.58592958,  14.1173234 , -15.21928743])
```

Problem 2 - Data Selection

In [67]:
```
n = 15

df = pd.read_excel(r'C:\Users\maste\Downloads\linear_regression_5e_data_sets\line
                   '\Appendices\data-table-B4.XLS')
sample = df.sample(n)
X = np.array(sample[['x4','x7','x9']])
y = np.array(sample[['y']])
```

```
WARNING *** OLE2 inconsistency: SSCS size is 0 but SSAT size is non-zero
```

In [70]:
```
sample[['y','x4','x7','x9']]
```

Out[70]:

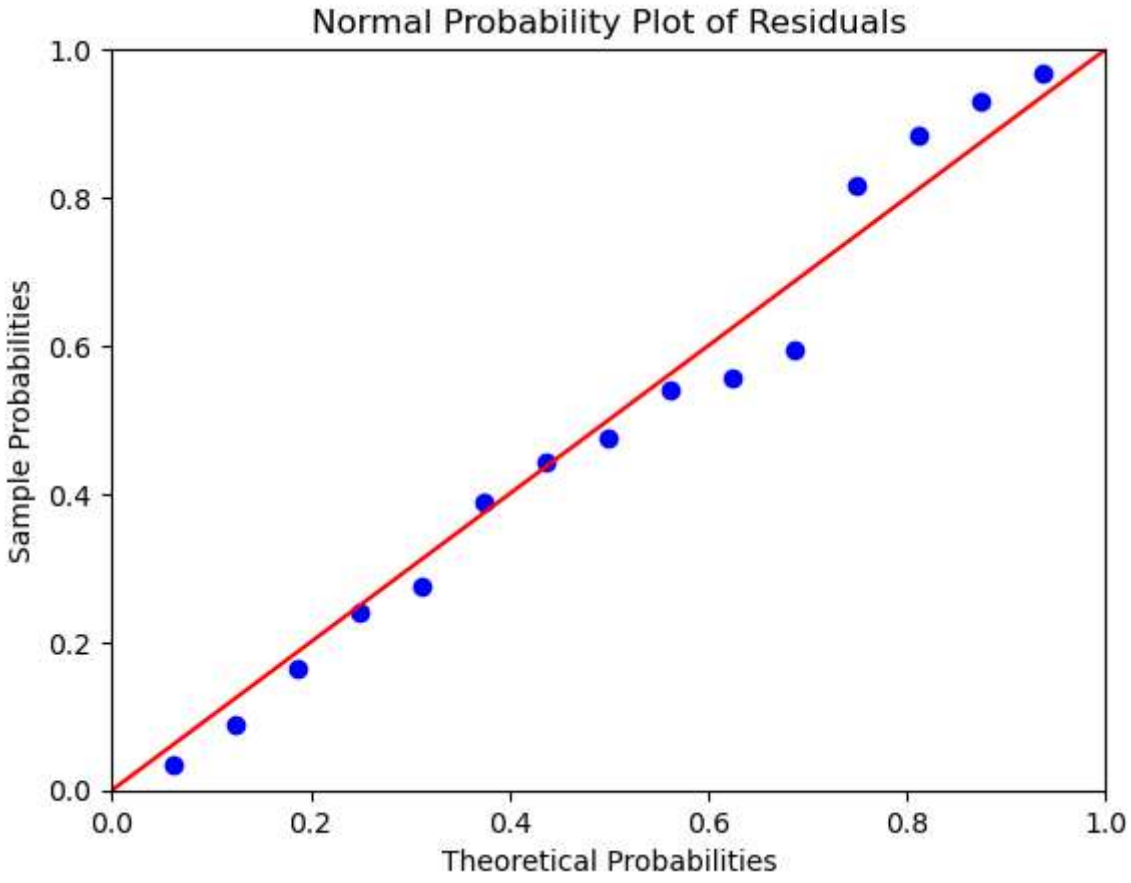
	y	x4	x7	x9
7	35.9	1.225	3	0
14	40.5	1.376	3	0
15	43.9	1.500	3	0
12	36.9	1.664	4	0
17	37.9	1.690	3	0
20	38.9	1.777	4	1
5	30.9	1.240	3	1
0	29.5	1.500	4	0
13	41.9	1.488	3	1
21	36.9	1.504	3	0
8	31.5	1.552	3	0
6	28.9	1.501	3	0
23	25.9	0.998	4	0
11	30.0	1.020	2	1
19	37.9	1.652	3	0

Problem 2(a)

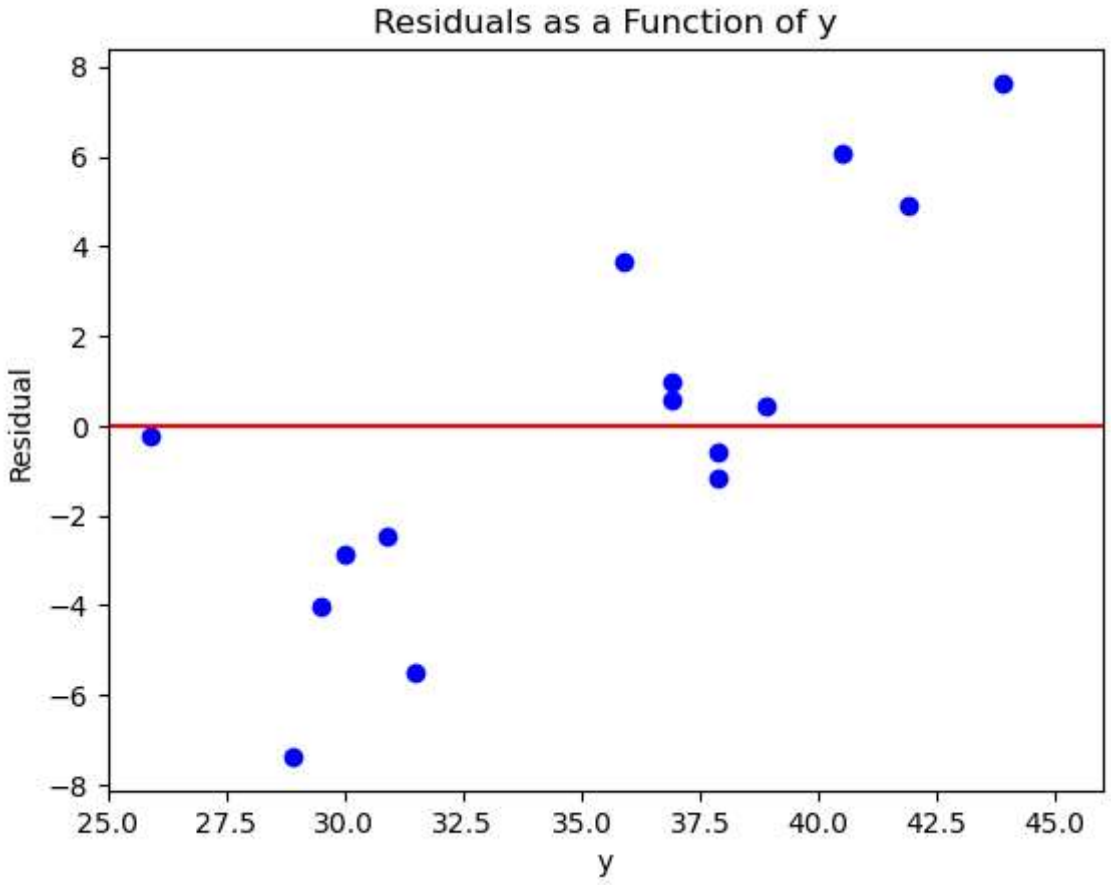
In [68]:
```
X = sm.add_constant(X)
mod = sm.OLS(y, X)
results = mod.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.374
Model:                            OLS   Adj. R-squared:                  0.203
Method:                 Least Squares   F-statistic:                     2.189
Date:                Tue, 15 Mar 2022   Prob (F-statistic):              0.147
Time:                        08:44:08   Log-Likelihood:                -42.419
No. Observations:                  15   AIC:                             92.84
Df Residuals:                      11   BIC:                             95.67
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         22.5038      9.954      2.261      0.045       0.595      44.412
x1            14.6679      5.805      2.527      0.028       1.890      27.445
x2            -2.7473      2.442     -1.125      0.285      -8.122       2.627
x3             0.9032      2.879      0.314      0.760      -5.434       7.240
==============================================================================
Omnibus:                        0.156   Durbin-Watson:                   1.084
Prob(Omnibus):                  0.925   Jarque-Bera (JB):                0.347
Skew:                           0.156   Prob(JB):                        0.841
Kurtosis:                       2.324   Cond. No.                         32.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.

C:\Users\maste\anaconda3\lib\site-packages\scipy\stats\stats.py:1603: UserWarni
ng: kurtosistest only valid for n>=20 ... continuing anyway, n=15
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

In [72]:
```
res = results.resid
pplot = sm.ProbPlot(res, stats.t, fit=True)
fig = pplot.ppplot(line="45")
h = plt.title("Normal Probability Plot of Residuals")
plt.show()
```



In [75]:
```
plt.scatter(y,res, c='b')
plt.plot([25,46],[0,0],c='r')
plt.xlim(25,46)
plt.xlabel('y')
plt.ylabel('Residual')
h = plt.title("Residuals as a Function of y")
plt.show()
```



Problem 3 - Data Selection

In [130]:
```
df = pd.DataFrame(np.array([[13.0,    4],
[34.2,    11],
[16.1,    5],
[65.6,    12],
[14.5,    6],
[49.2,    13],
[17.8,    7],
[66.2,    14],
[22.0,    8],
[81.2,    15],
[27.4,    9],
[87.4,    16],
[16.8,    10],
[114.5,    17]]), columns=['y','x'])
```

In [131]:
```
n = 14

sample = df.sample(n)
X = np.array(sample[['x']])
y = np.array(sample[['y']])
```

In [132]:
```
sample
```

Out[132]:

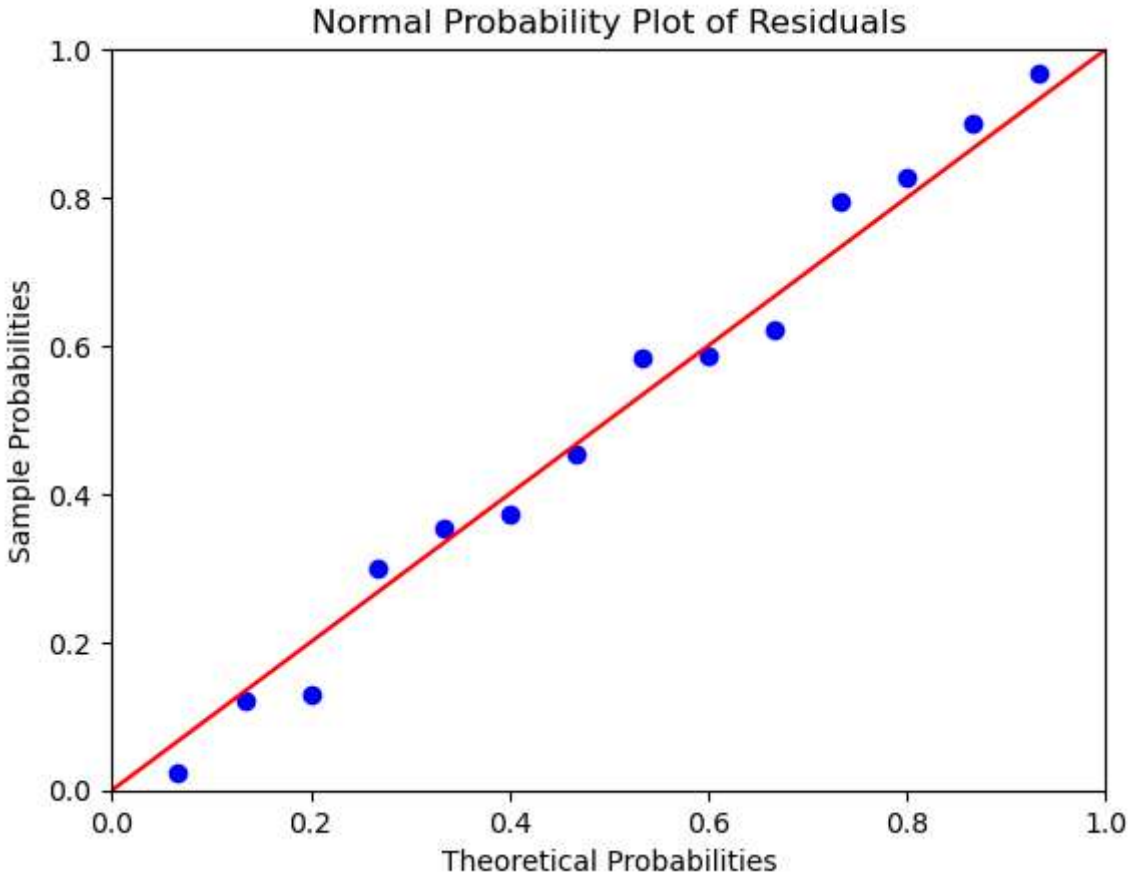
	y	x
9	81.2	15.0
0	13.0	4.0
12	16.8	10.0
3	65.6	12.0
10	27.4	9.0
8	22.0	8.0
7	66.2	14.0
4	14.5	6.0
6	17.8	7.0
2	16.1	5.0
13	114.5	17.0
11	87.4	16.0
1	34.2	11.0
5	49.2	13.0

Problem 3(a)

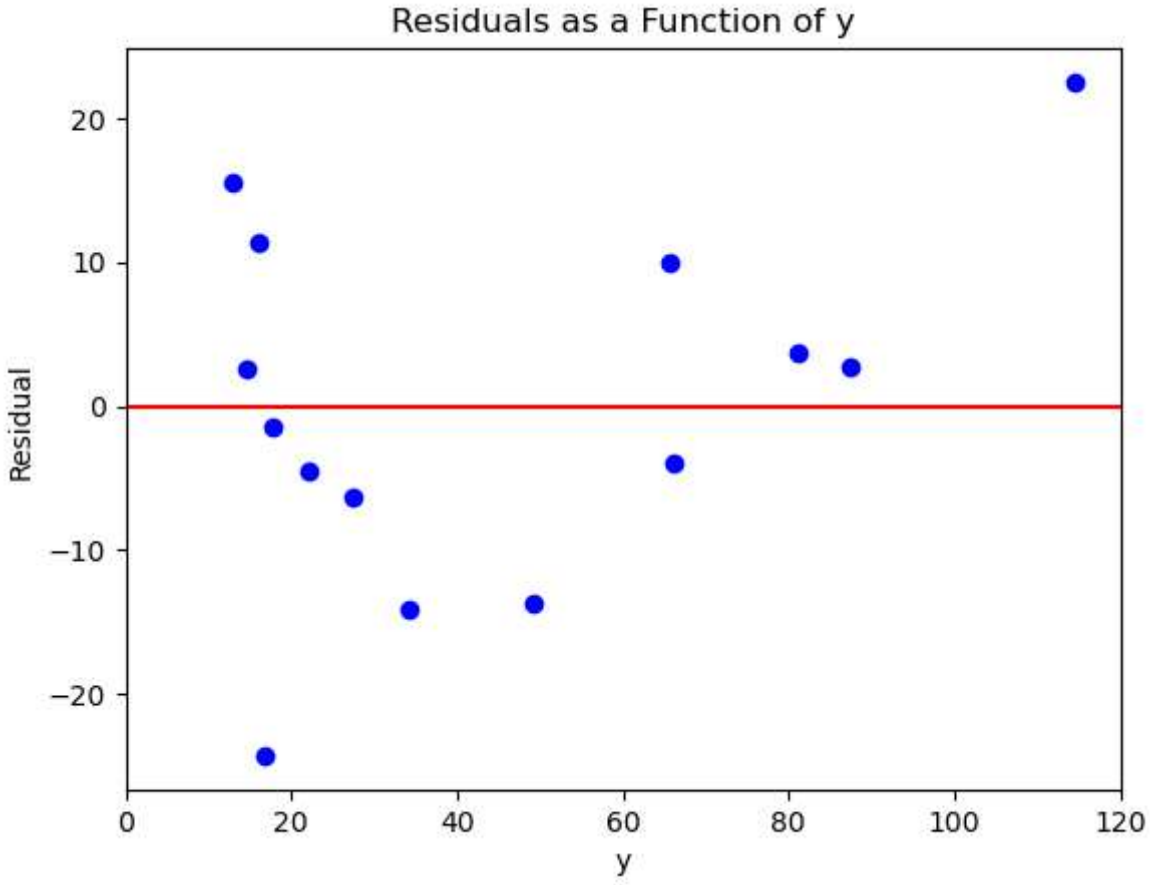
In [133]:
```
X = sm.add_constant(X)
mod = sm.OLS(y, X)
results = mod.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.854
Model:                            OLS   Adj. R-squared:                  0.842
Method:                 Least Squares   F-statistic:                     70.09
Date:                Tue, 15 Mar 2022   Prob (F-statistic):           2.35e-06
Time:                        12:12:31   Log-Likelihood:                -54.813
No. Observations:                  14   AIC:                             113.6
Df Residuals:                      12   BIC:                             114.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -31.6982      9.776     -3.243      0.007     -52.998     -10.399
x1             7.2767      0.869      8.372      0.000       5.383       9.170
==============================================================================
Omnibus:                        0.049   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.976   Jarque-Bera (JB):                0.153
Skew:                          -0.095   Prob(JB):                        0.926
Kurtosis:                       2.524   Cond. No.                         31.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.

C:\Users\maste\anaconda3\lib\site-packages\scipy\stats\stats.py:1603: UserWarni
ng: kurtosistest only valid for n>=20 ... continuing anyway, n=14
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

In [134]:
```
res = results.resid
pplot = sm.ProbPlot(res, stats.t, fit=True)
fig = pplot.ppplot(line="45")
h = plt.title("Normal Probability Plot of Residuals")
plt.show()
```



In [135]:
```
plt.scatter(y,res, c='b')
plt.plot([0,120],[0,0],c='r')
plt.xlim(0,120)
plt.xlabel('y')
plt.ylabel('Residual')
h = plt.title("Residuals as a Function of y")
plt.show()
```

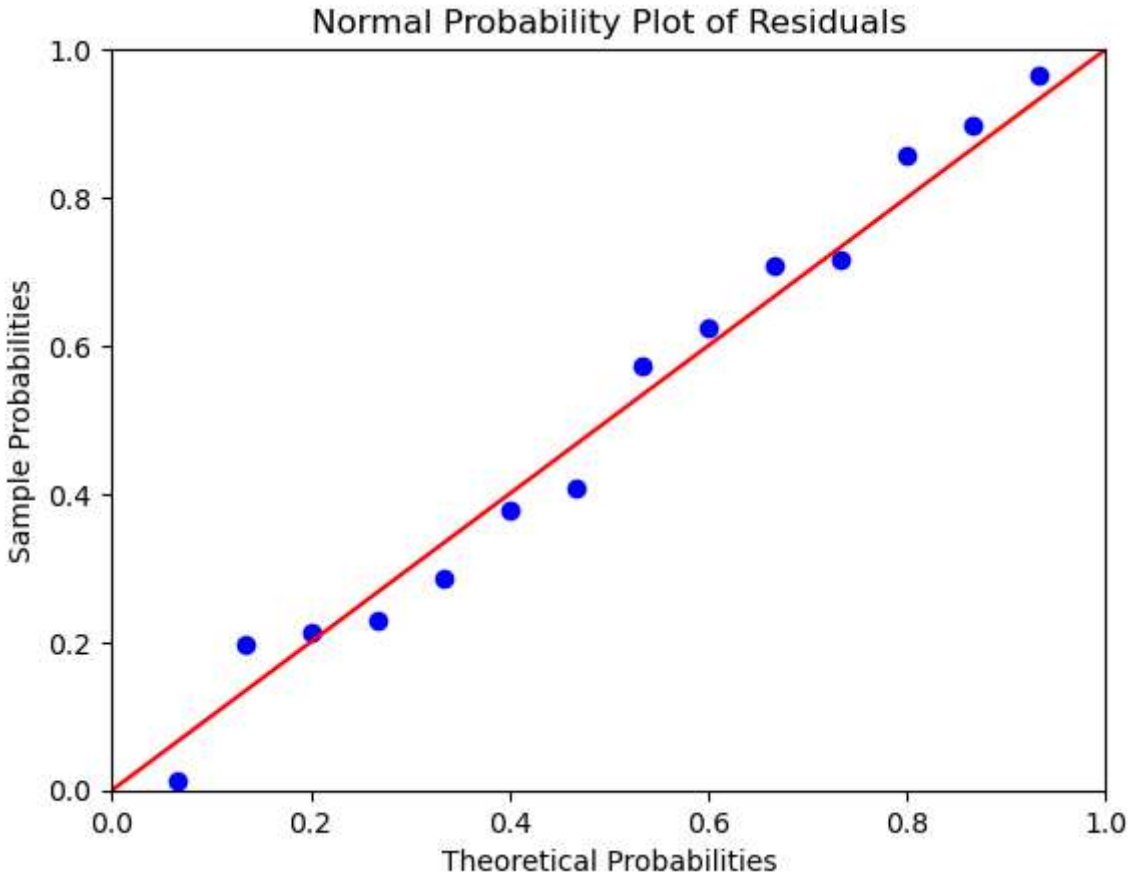


Problem 3(b)

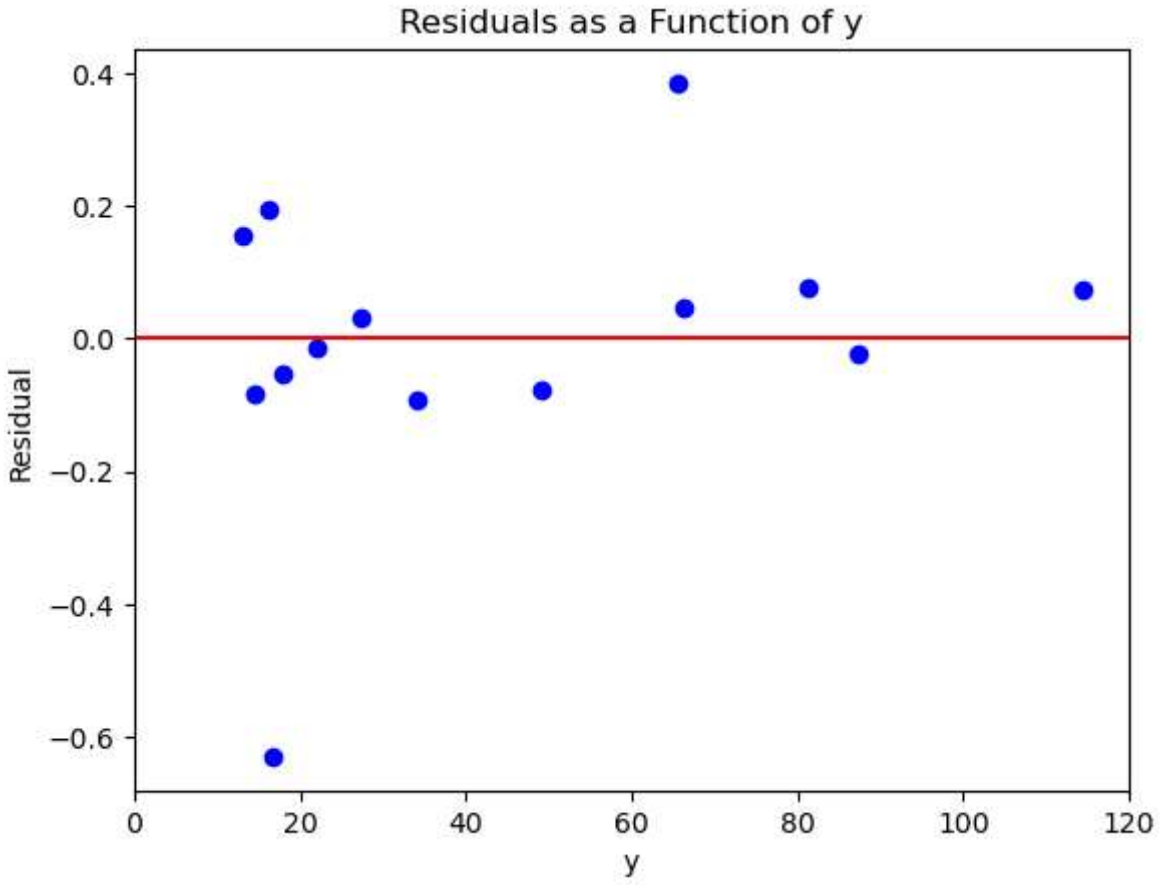
In [136]:
```
X = sm.add_constant(X)
mod = sm.OLS(np.log(y), X)
results = mod.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.914
Model:                            OLS   Adj. R-squared:                  0.907
Method:                 Least Squares   F-statistic:                     127.1
Date:                Tue, 15 Mar 2022   Prob (F-statistic):           9.68e-08
Time:                        12:12:33   Log-Likelihood:                 1.6588
No. Observations:                  14   AIC:                            0.6825
Df Residuals:                      12   BIC:                             1.961
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.7162      0.173      9.914      0.000       1.339       2.093
x1             0.1735      0.015     11.273      0.000       0.140       0.207
==============================================================================
Omnibus:                       12.871   Durbin-Watson:                   2.922
Prob(Omnibus):                  0.002   Jarque-Bera (JB):                9.564
Skew:                          -1.313   Prob(JB):                      0.00838
Kurtosis:                       6.082   Cond. No.                         31.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.

C:\Users\maste\anaconda3\lib\site-packages\scipy\stats\stats.py:1603: UserWarni
ng: kurtosistest only valid for n>=20 ... continuing anyway, n=14
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

In [137]:
```
res = results.resid
pplot = sm.ProbPlot(res, stats.t, fit=True)
fig = pplot.ppplot(line="45")
h = plt.title("Normal Probability Plot of Residuals")
plt.show()
```



In [138]:
```
plt.scatter(y,res, c='b')
plt.plot([0,120],[0,0],c='r')
plt.xlim(0,120)
plt.xlabel('y')
plt.ylabel('Residual')
h = plt.title("Residuals as a Function of y")
plt.show()
```



In [ ]:
```
```