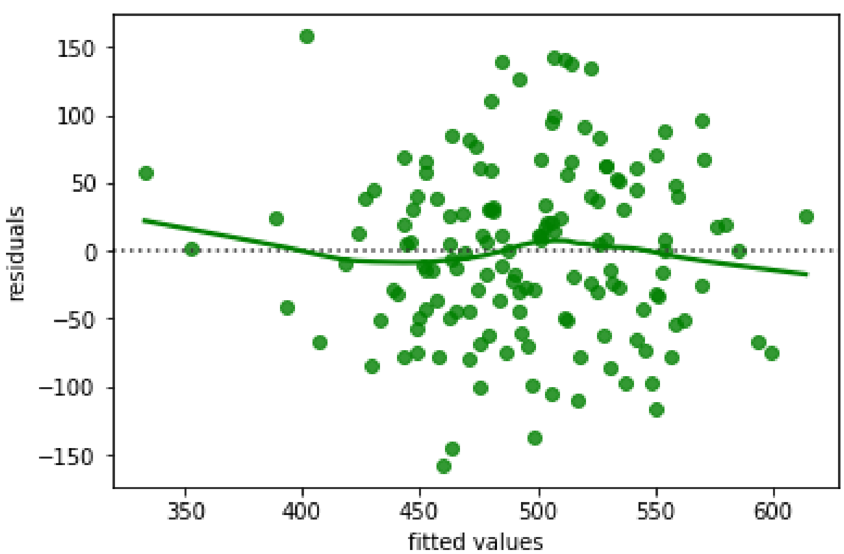
7.2) Fit the regression plane for the fathers using FFVC as the dependent variable and age and height as the independent variables.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

dataset = pd.read_excel('Lung Function.xls')
X = dataset.iloc[:, 3:5].values
y = dataset.iloc[:, 6].values

import statsmodels.api as sm
X = np.append(arr = np.ones((150, 1)).astype(int), values = X, axis = 1)

regressor_OLS = sm.OLS(endog = y, exog = X).fit()

y_pred = regressor_OLS.predict(X)

regressor_OLS.summary()

model_norm_residuals = regressor_OLS.get_influence().resid_studentized_internal


fig1=sns.residplot(y_pred,y, lowess=True, color="green")
fig1.set(xlabel='fitted values', ylabel='residuals')
plt.show()

fig = sm.qqplot(model_norm_residuals,color="blue")
plt.show()
```



There is a Linear relationship between FFVC, age and height. We observe normal distribution between dependent and independent variables. There are also some outliers

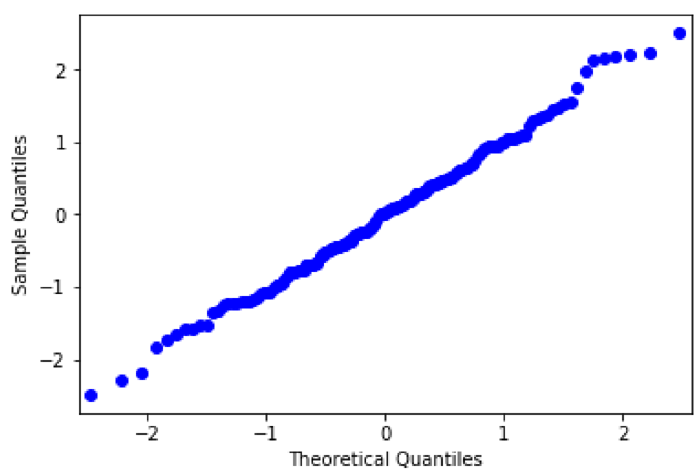
7.3) Write the results for Problem 7.2 so they would be suitable for inclusion in a report. Include table(s) that present the results the reader should see.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.360
Model:                            OLS   Adj. R-squared:                  0.352
Method:                 Least Squares   F-statistic:                     41.40
Date:                Thu, 27 Sep 2018   Prob (F-statistic):           5.49e-15
Time:                        13:21:28   Log-Likelihood:                -834.94
No. Observations:                 150   AIC:                             1676.
Df Residuals:                     147   BIC:                             1685.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -453.9204    135.965     -3.338      0.001    -722.620    -185.221
x1            -2.7788      0.761     -3.651      0.000      -4.283      -1.275
x2            15.3144      1.887      8.116      0.000      11.586      19.043
==============================================================================
Omnibus:                        0.669   Durbin-Watson:                   2.056
Prob(Omnibus):                  0.716   Jarque-Bera (JB):                0.799
Skew:                           0.133   Prob(JB):                        0.671
Kurtosis:                       2.762   Cond. No.                     2.09e+03
==============================================================================
```

So, as per the values in the above mentioned table the equation is :

FFVC=-453.9204 – 2.7788 * (Age Of The Father) + 15.3144 * (Height Of The Father)

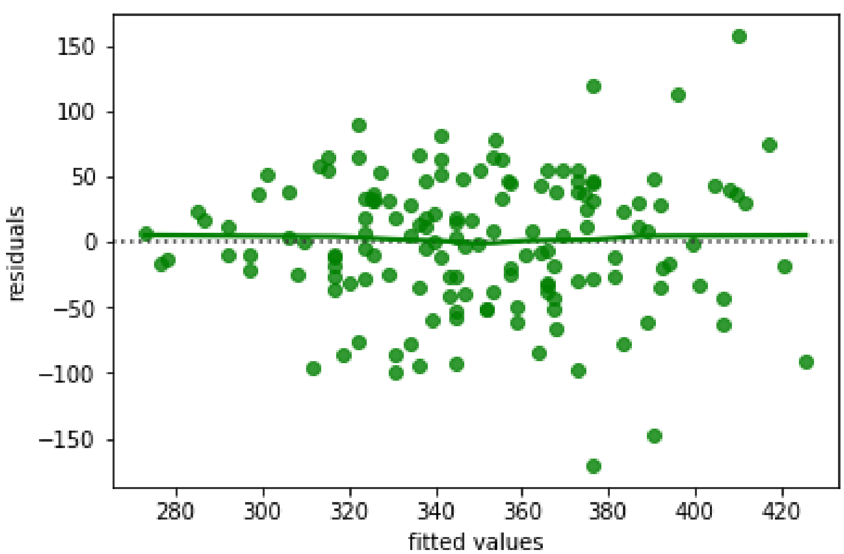
QQPLOT:



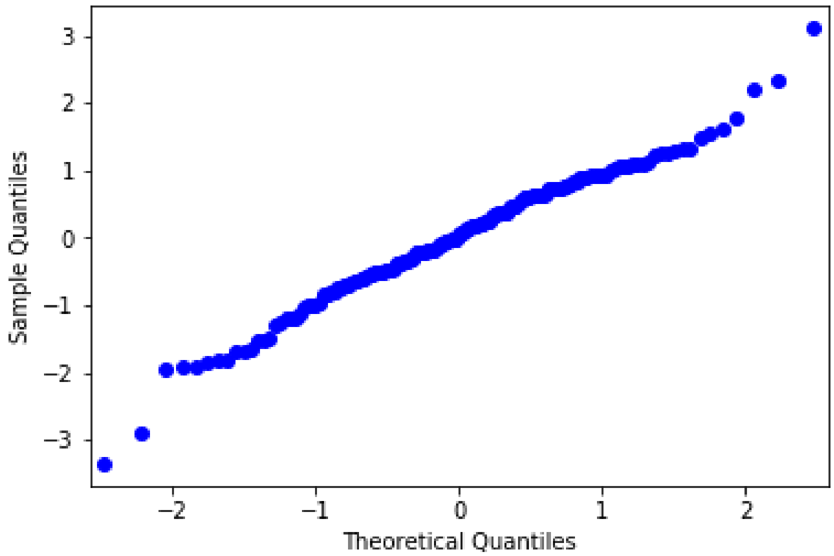
R-sqaured value is 0.360 =>36% variation is explained by age and height of father

7.4) Fit the regression plane for mothers with MFVC as the dependent variable and age and height as the independent variables. Summarize the results in a tabular form. Test whether the regression results for mothers and fathers are significantly different.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

dataset = pd.read_excel('Lung Function.xls')
X = dataset.iloc[:, 9:11].values
y = dataset.iloc[:, 12].values

import statsmodels.api as sm
X = np.append(arr = np.ones((150, 1)).astype(int), values = X, axis = 1)

regressor_OLS = sm.OLS(endog = y, exog = X).fit()

y_pred = regressor_OLS.predict(X)

regressor_OLS.summary()

model_norm_residuals = regressor_OLS.get_influence().resid_studentized_internal


fig1=sns.residplot(y_pred,y, lowess=True, color="green")
fig1.set(xlabel='fitted values', ylabel='residuals')
plt.show()

fig = sm.qqplot(model_norm_residuals,color="blue")
plt.show()
```



There is a Linear relationship between MFVC, age and height. We observe normal distribution between dependent and independent variables. There are also some outliers
QQPLOT:



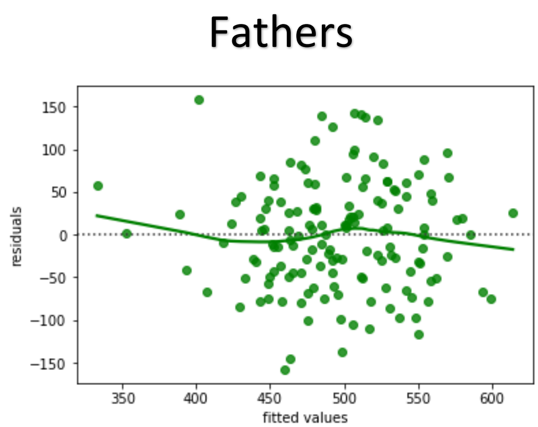

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.288
Model:                            OLS   Adj. R-squared:                  0.279
Method:                 Least Squares   F-statistic:                     29.76
Date:                Thu, 27 Sep 2018   Prob (F-statistic):           1.41e-11
Time:                        13:31:44   Log-Likelihood:                -802.06
No. Observations:                 150   AIC:                             1610.
Df Residuals:                     147   BIC:                             1619.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -372.0288    111.349     -3.341      0.001    -592.081    -151.977
x1            -1.7681      0.626     -2.823      0.005      -3.006      -0.530
x2            12.3050      1.703      7.226      0.000       8.940      15.670
==============================================================================
Omnibus:                        4.670   Durbin-Watson:                   2.138
Prob(Omnibus):                  0.097   Jarque-Bera (JB):                4.533
Skew:                          -0.282   Prob(JB):                        0.104
Kurtosis:                       3.639   Cond. No.                     1.98e+03
==============================================================================
```

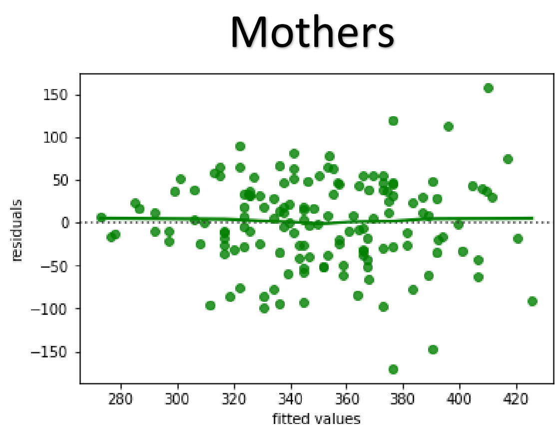
The equation:

MFVC=-372.0288 – 1.7681 * (Age Of The Mother) + 12.3050 * (Height Of The Mother)

R-sqaured value is 0.288 =>28.8% variation is explained by age and height of Mother

Comparing residual plots for Fathers and Mothers , Height and Age being the independent variables:





In the mothers graph the residual line is very close to zero.

From the graph we can state that there is a better linear relationship observed between independent and dependent variables for mothers.

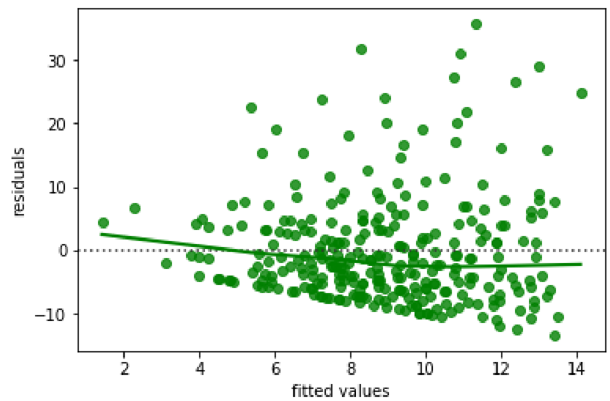
7.5) From the depression data set described in Table 3.4, predict the reported level of depression as given by CESD, using INCOME, SEX, and AGE as independent variables. Analyze the residuals and decide whether or not it is reasonable to assume that they follow a normal distribution.

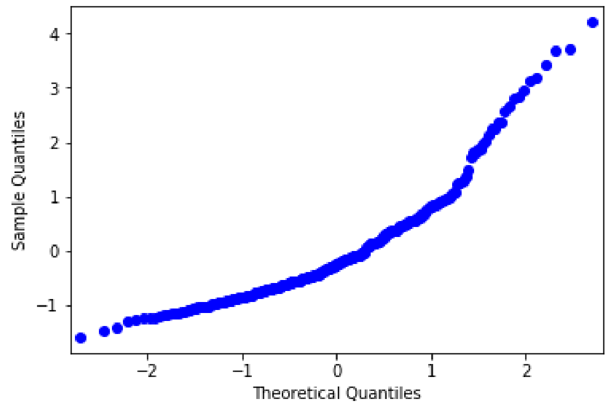
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

dataset = pd.read_excel('Depression.xls')
X = dataset.iloc[:,[1,2,6]].values
y = dataset.iloc[:, -9].values

import statsmodels.api as sm
X = np.append(arr = np.ones((294, 1)).astype(int), values = X, axis = 1)

regressor_OLS = sm.OLS(endog = y, exog = X).fit()

y_pred = regressor_OLS.predict(X)

regressor_OLS.summary()

model_norm_residuals = regressor_OLS.get_influence().resid_studentized_internal


fig1=sns.residplot(y_pred,y, lowess=True, color="green")
fig1.set(xlabel='fitted values', ylabel='residuals')
plt.show()

fig = sm.qqplot(model_norm_residuals,color="blue")
plt.show()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.064
Method:                 Least Squares   F-statistic:                     7.718
Date:                Thu, 27 Sep 2018   Prob (F-statistic):           5.61e-05
Time:                        13:48:13   Log-Likelihood:                -1045.5
No. Observations:                 294   AIC:                             2099.
Df Residuals:                     290   BIC:                             2114.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         12.4490      2.419      5.146      0.000       7.687      17.211
x1             1.8203      1.044      1.744      0.082      -0.234       3.875
x2            -0.0989      0.028     -3.522      0.000      -0.154      -0.044
x3            -0.1032      0.034     -3.058      0.002      -0.170      -0.037
==============================================================================
Omnibus:                       90.599   Durbin-Watson:                   1.671
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              202.999
Skew:                           1.525   Prob(JB):                     8.30e-45
Kurtosis:                       5.696   Cond. No.                         265.
==============================================================================
```

Equation:

CESD = 12.4490 + 1.8203*SEX - 0.0989*AGE– 0.1032*INCOME





From the above residual plot we can observe that many points lie away from residual line. There are many outliers. We can also observe a decreasing linear relation between dependent and independent variables. Therefore, we can infer that the data is not normally distributed.

8.1) Use the depression data set described in Table 3.4. Using CESD as the dependent variable, and age, income, and level of education as the independent variables, run a forward stepwise regression program to determine which of the independent variables predict level of depression for women.

(done using R)

<u>CODE:</u>

```
depression<-read.csv(file.choose(),header=TRUE)
View(depression)

depression1<-subset(depression,SEX=="2")
View(depression1)
attach(depression1)

tstep<-step(lm(CESD~1,depression1),direction="forward",scope=~AGE+INCOME+EDUCA
T)
```

```
Start:          AIC=828.09
CESD ~ 1
Df Sum of Sq    RSS       AIC
+ AGE           1         497.32 16211 824.56
<none>                           16708 828.09
+ INCOME        1         180.40 16528 828.10
+ EDUCAT        1         107.07 16601 828.91
```

```
Step:   AIC=824.56
CESD ~ AGE
Df Sum of Sq    RSS     AIC
+ INCOME        1       334.61 15876 822.74
+ EDUCAT        1       236.61 15974 823.87
<none>                         16211 824.56
Step:   AIC=822.74
CESD ~ AGE + INCOME
Df Sum of Sq    RSS     AIC
<none>                         15876 822.74
+ EDUCAT        1       100.8 15775 823.58

summary(tstep)

Call:

lm(formula = CESD ~ AGE + INCOME, data = depression1)

Residuals:

Min     1Q      Median  3Q      Max
-13.440 -7.004  -2.532  4.528   35.477
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.22231    2.22641 7.286 9.65e-12  ***
AGE     -0.10445        0.03843 -2.718  0.00721 **
INCOME -0.09696         0.04978 -1.948  0.05300 .
---

Signif. codes:    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.391 on 180 degrees of freedom

Multiple R-squared:     0.04979, Adjusted R-squared:     0.03923

F-statistic: 4.716 on 2 and 180 DF,   p-value: 0.01008

Detach(depression1)
```

Regression equation:

Depression for woman= 16.22-0.104*AGE-0.097*INCOME

8.5 Using the data given in Table 8.1, repeat the analyses described in this chapter with (P/E) 1/2 as the dependent variable instead of P/E. Do the results change much? Does it make sense to use the square root transformation?

(done in R)

CODE:

```
chemical<-read.csv(file.choose(),header=TRUE)
View(chemical)
attach(chemical)

rootPE<-sqrt(pe)
View(rootPE)

step(lm(pe~1,chemical),scope=~ror5+de+salesgr5+eps5+npm1+payoutr1)

Start:  AIC=62.71
pe ~ 1
Df Sum of Sq      RSS     AIC
+ de    1       50.889 176.08       57.092
+ npm1         1       28.012 198.96       60.756
+ payoutr1   1       24.637 202.33       61.261
+ ror5 1       22.619 204.35       61.559

<none>                                  226.97 62.708
+ eps5                 1       8.140  218.83 63.613
+ salesgr5     1       3.854  223.11 64.194
Step:   AIC=57.09
pe ~ de
Df Sum of Sq RSS     AIC
+ payoutr1   1       14.868 161.21       56.445
+ npm1                 1       14.328 161.75       56.546
<none>                                  176.08         57.092
+ salesgr5     1       3.347 172.73          58.516
+ ror5          1       2.816 173.26          58.608
+ eps5                 1       2.283 173.79          58.700
- de            1       50.889 226.97       62.708
Step:   AIC=56.45
pe ~ de + payoutr1
Df Sum of Sq RSS     AIC
+ npm1                 1       42.026 119.18       49.384
+ salesgr5     1       16.333 144.88       55.241
```

```
+ ror5       1   13.415 147.79   55.839
<none>                           161.21   56.445
- payoutr1   1   14.868 176.08 57.092
+ eps5           1      0.602 160.61    58.333
- de         1   41.120 202.33   61.261
Step:  AIC=49.38
pe ~ de + payoutr1 + npm1
Df Sum of Sq RSS    AIC
+ salesgr5   1   14.240     104.94 47.567
<none>                           119.18 49.384
- de         1   15.222     134.41 50.990
+ ror5       1   0.339  118.84 51.299
+ eps5           1      0.257  118.93 51.319
- npm1           1      42.026      161.21 56.445
- payoutr1   1   42.566     161.75 56.546
Step:  AIC=47.57

pe ~ de + payoutr1 + npm1 + salesgr5

Df Sum of Sq        RSS    AIC
<none>                      104.94     47.567
- de    1     12.738     117.68     49.004
+ eps5       1   1.619  103.33     49.101
- salesgr5   1   14.240     119.18     49.384
+ ror5 1     0.189  104.75     49.513
- npm1       1   39.933     144.88     55.241
- payoutr1   1   56.008     160.95     58.397
Call:

lm(formula = pe ~ de + payoutr1 + npm1 + salesgr5, data = chemical)

Coefficients:

(Intercept)   de      payoutr1      npm1 salesgr5

1.2771        -3.1609       10.7490      0.3523        0.1949
```

P/E = -3.1609(D/E) + 10.7490(PAYOUTR1)+ 0.3523(NPM1)+ 0.1949(SALESGR5) + 1.2771

```
step(lm(rootPE~1,chemical),scope=~ror5+de+salesgr5+eps5+npm1+payoutr1)
```

```
Start:  AIC=-45.47
rootPE ~ 1
Df Sum of Sq RSS AIC
+ de 1 1.39216 4.7735 -51.144
+ payoutr1 1 0.70597 5.4597 -47.114
+ npm1 1 0.66516 5.5005 -46.891
+ ror5 1 0.53346 5.6322 -46.181
<none> 6.1656 -45.466
+ eps5 1 0.28401 5.8816 -44.881
+ salesgr5 1 0.11076 6.0549 -44.010
Step:  AIC=-51.14
rootPE ~ de
Df Sum of Sq RSS AIC
+ payoutr1 1 0.43202 4.3414 -51.990
+ npm1 1 0.31886 4.4546 -51.218
<none> 4.7735 -51.144
+ eps5 1 0.09741 4.6761 -49.762
+ salesgr5 1 0.09654 4.6769 -49.757
+ ror5 1 0.04592 4.7275 -49.434
- de 1 1.39216 6.1656 -45.466
Step:  AIC=-51.99
rootPE ~ de + payoutr1
Df Sum of Sq RSS AIC
+ npm1 1 1.02581 3.3156 -58.076
+ salesgr5 1 0.47278 3.8687 -53.449
+ ror5 1 0.29712 4.0443 -52.117
<none> 4.3414 -51.990
- payoutr1 1 0.43202 4.7735 -51.144
+ eps5 1 0.00443 4.3370 -50.020
- de 1 1.11822 5.4597 -47.114
Step:  AIC=-58.08
rootPE ~ de + payoutr1 + npm1
Df Sum of Sq RSS AIC
+ salesgr5 1 0.41705 2.8986 -60.109
<none> 3.3156 -58.076
- de 1 0.43827 3.7539 -56.352
+ eps5 1 0.01824 3.2974 -56.242
+ ror5 1 0.00360 3.3120 -56.109
- npm1 1 1.02581 4.3414 -51.990
- payoutr1 1 1.13897 4.4546 -51.218

Step:  AIC=-60.11
```

rootPE ~ de + payoutr1 + npm1 + salesgr5

	Df	Sum of Sq	RSS	AIC
<none>			2.8986	-60.109
+ eps5	1	0.07126	2.8273	-58.856
- de	1	0.36619	3.2648	-58.540
+ ror5	1	0.01336	2.8852	-58.248
- salesgr5	1	0.41705	3.3156	-58.076
- npm1	1	0.97008	3.8687	-53.449
- payoutr1	1	1.52633	4.4249	-49.418

Call:

lm(formula = rootPE ~ de + payoutr1 + npm1 + salesgr5, data = chemical)

Coefficients:

(Intercept)	de	payoutr1	npm1	salesgr5
1.70082	-0.53592	1.77446	0.05491	0.03335

(P/E)^(1/2) = -0.53592(D/E) + 1.77446(PAYOUTR1)+ 0.05491(NPM1)+ 0.03335(SALESGR5) + 1.70082

So , We can state that there is not much difference in results after taking the square root between the two models.