In [274]:

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import pylab

#读取数据
df=pd.read_csv('horse-colic.data.txt',sep=' ')
attrname=['surgery','Age','Hospital_Number','rectal_temperature','pulse','respirato
labelattr=[1,2,3,7,8,9,10,11,12,13,14,15,17,18,21,23,24,25,26,27,28]
valueattr=[4,5,6,16,19,20,22]
```

In [196]:

```
#分析标签属性的出现频数
for attrid in labelattr:
    print(df[attrname[attrid-1]].value_counts())
    print()
```

```
1    180
2    119
?      1
Name: surgery, dtype: int64

1    276
9     24
Name: Age, dtype: int64

529796     2
529424     2
5279822    2
527916     2
530526     2
528996     2
528729     2
528151     2
529461     2
528931     2
528890     2
528469     2
530693     2
532349     2
528904     2
527544     2
533697     1
533696     1
533736     1
534719     1
530101     1
529399     1
533692     1
528570     1
527957     1
534157     1
521399     1
529045     1
530612     1
528047     1
          ..
533887     1
535031     1
530301     1
5294369    1
530297     1
529272     1
535415     1
530294     1
534899     1
529777     1
535407     1
522979     1
528743     1
534885     1
```

```
534857     1
530276     1
535392     1
535338     1
527709     1
527706     1
533847     1
528214     1
535381     1
533750     1
527698     1
530255     1
530254     1
533836     1
530251     1
535043     1
Name: Hospital_Number, dtype: int64

3    109
1     78
?     56
2     30
4     27
Name: temperature_of_extremities, dtype: int64

1    115
3    103
?     69
4      8
2      5
Name: peripheral_pulse, dtype: int64

1    79
3    58
?    47
4    41
2    30
5    25
6    20
Name: mucous_membranes, dtype: int64

1    188
2     78
?     32
3      2
Name: capillary_refill_time, dtype: int64

3    67
2    59
?    55
5    42
4    39
1    38
Name: pain, dtype: int64

3    128
4     73
?     44
1     39
2     16
Name: peristalsis, dtype: int64
```

```
1    76
2    65
3    65
?    56
4    38
Name: abdominal_distension, dtype: int64

?    104
2    102
1     71
3     23
Name: nasogastric_tube, dtype: int64

1    120
?    106
3     39
2     35
Name: nasogastric_reflux, dtype: int64

?    102
4     79
1     57
3     49
2     13
Name: rectal_examination, dtype: int64

?    118
5     79
4     43
1     28
2     19
3     13
Name: abdomen, dtype: int64

?    165
2     48
3     46
1     41
Name: abdominocentesis_appearance, dtype: int64

1    178
2     77
3     44
?      1
Name: outcome, dtype: int64

1    191
2    109
Name: surgical_lesion, dtype: int64

0        56
3111     33
3205     29
2208     20
2205     13
4205     11
2209     11
2124      9
1400      8
31110     7
7111      7
```

```
2113      6
2112      5
400       5
3209      4
4300      4
2206      4
5400      4
3112      3
4124      3
2111      3
2207      3
7209      3
5206      2
5124      2
3124      2
5111      2
9400      2
6111      2
2322      2
         ..
3025      2
8400      2
6112      2
11300     1
4122      1
7113      1
6209      1
3115      1
5000      1
3133      1
4111      1
3400      1
300       1
12208     1
9000      1
5205      1
1111      1
1124      1
8300      1
2305      1
4206      1
4207      1
21110     1
2300      1
3207      1
11400     1
7400      1
3113      1
3300      1
41110     1
Name: #1_lesion, dtype: int64

0       293
3111      3
6112      1
7111      1
1400      1
3112      1
Name: #2_lesion, dtype: int64

0       299
```

```
2209      1
Name: #3_lesion, dtype: int64

2    201
1     99
Name: cp_data, dtype: int64
```

In [263]:

```
#数值属性的最小值，1/4分位数，中位数，均值，3/4分位数，最大值
for attrid in valueattr:
    print(attrname[attrid - 1])
    series=df[attrname[attrid - 1]].apply(pd.to_numeric, errors='coerce')
    series=series[series.notnull()]
    print('min:',series.min())
    print('1/4 quantile:',series.quantile(0.25))
    print('mean:',series.mean())
    print('median:',series.median())
    print('3/4 quantile:',series.quantile(0.75))
    print('max:',series.max())
    print()
```

```
rectal_temperature
min: 35.4
1/4 quantile: 37.8
mean: 38.16791666666669
median: 38.2
3/4 quantile: 38.5
max: 40.8

pulse
min: 30.0
1/4 quantile: 48.0
mean: 71.91304347826087
median: 64.0
3/4 quantile: 88.0
max: 184.0

respiratory_rate
min: 8.0
1/4 quantile: 18.5
mean: 30.417355371900825
median: 24.5
3/4 quantile: 36.0
max: 96.0

nasogastric_reflux_PH
min: 1.0
1/4 quantile: 3.0
mean: 4.707547169811321
median: 5.0
3/4 quantile: 6.5
max: 7.5

packed_cell_volume
min: 23.0
1/4 quantile: 38.0
mean: 46.29520295202952
median: 45.0
3/4 quantile: 52.0
max: 75.0

total_protein
min: 3.3
1/4 quantile: 6.5
mean: 24.456928838951317
median: 7.5
3/4 quantile: 57.0
```

```
max: 89.0

abdomcentesis_total_protein
min: 0.1
1/4 quantile: 2.0
mean: 3.0196078431372553
median: 2.25
3/4 quantile: 3.9
max: 10.1
```

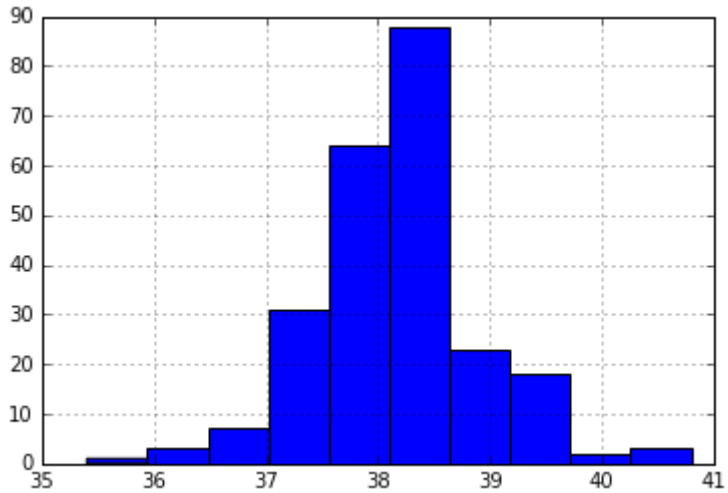
In [203]:

```
#直方图
attrid=4
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
series.hist()
```

```
rectal_temperature
```

Out[203]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2507e6eb70>
```



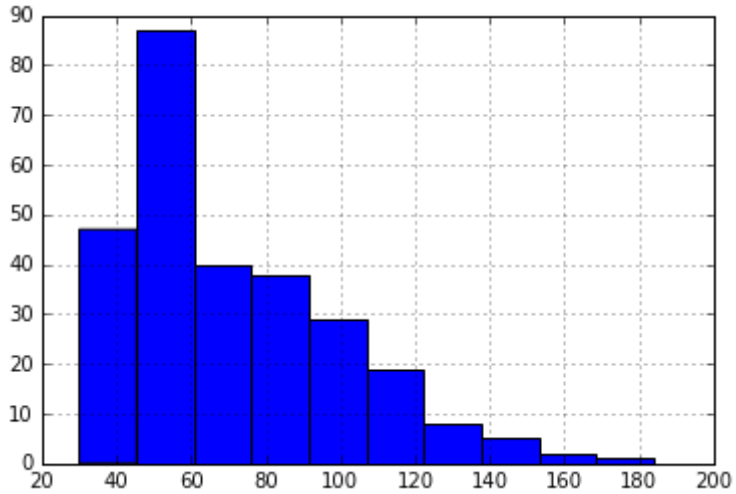
In [204]:

```
#直方图
attrid=5
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
series.hist()
```

```
pulse
```

Out[204]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2507df9128>
```



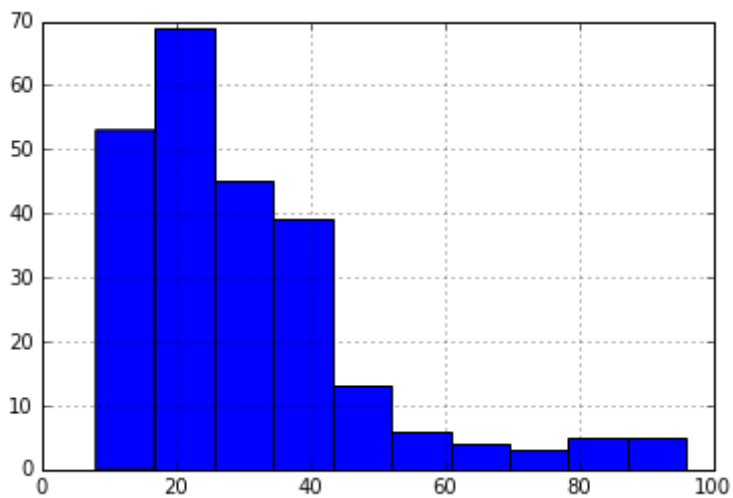
In [205]:

```
#直方图
attrid=6
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
series.hist()
```

```
respiratory_rate
```

Out[205]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2507d03400>
```



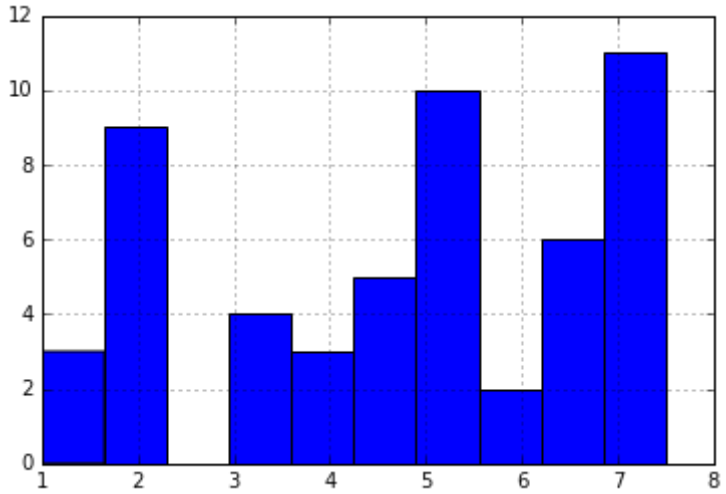
In [206]:

```
#直方图
attrid=16
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
series.hist()
```

```
nasogastric_reflux_PH
```

Out[206]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2507ca0748>
```



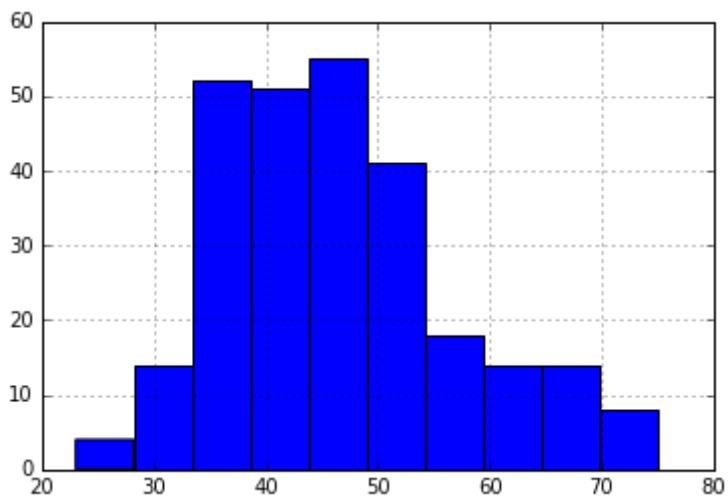
In [207]:

```
#直方图
attrid=19
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
series.hist()
```

```
packed_cell_volume
```

Out[207]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2507c26240>
```



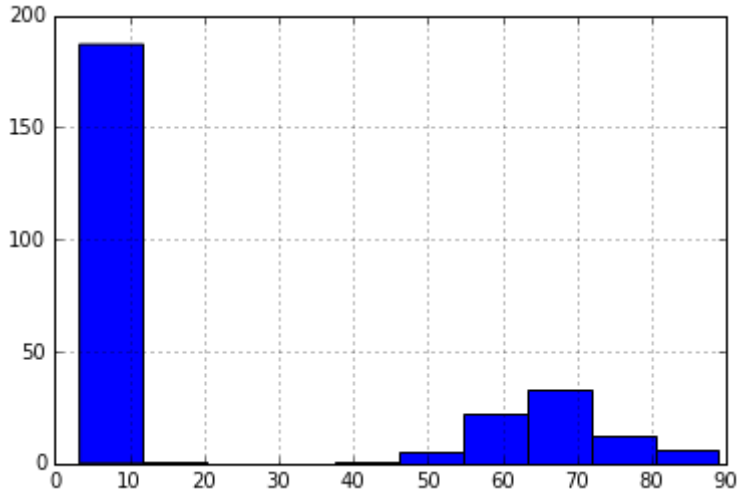
In [208]:

```
#直方图
attrid=20
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
series.hist()
```

total_protein

Out[208]:

<matplotlib.axes._subplots.AxesSubplot at 0x7f2507c16ef0>



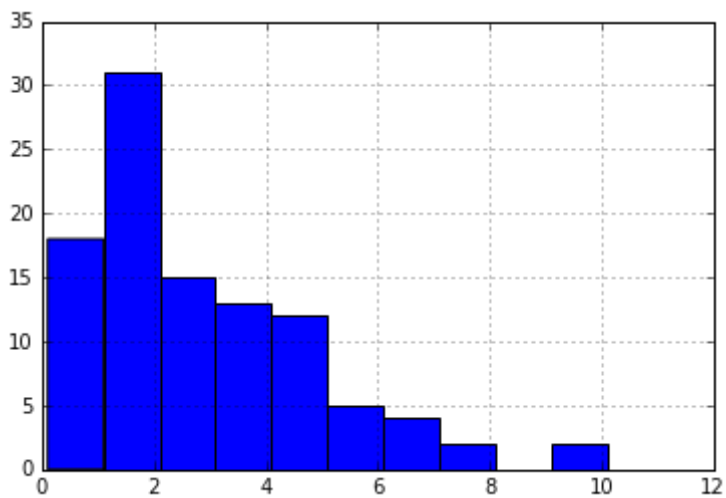
In [209]:

```
#直方图
attrid=22
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
series.hist()
```

abdomcentesis_total_protein

Out[209]:

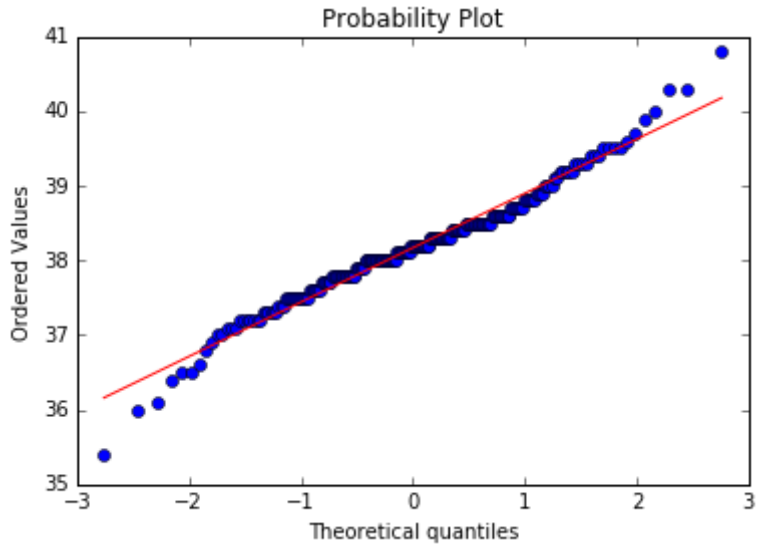
<matplotlib.axes._subplots.AxesSubplot at 0x7f2507d7b550>



In [231]:

```
#qq图
attrid=4
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=stats.probplot(series, dist="norm", plot=pylab)
```

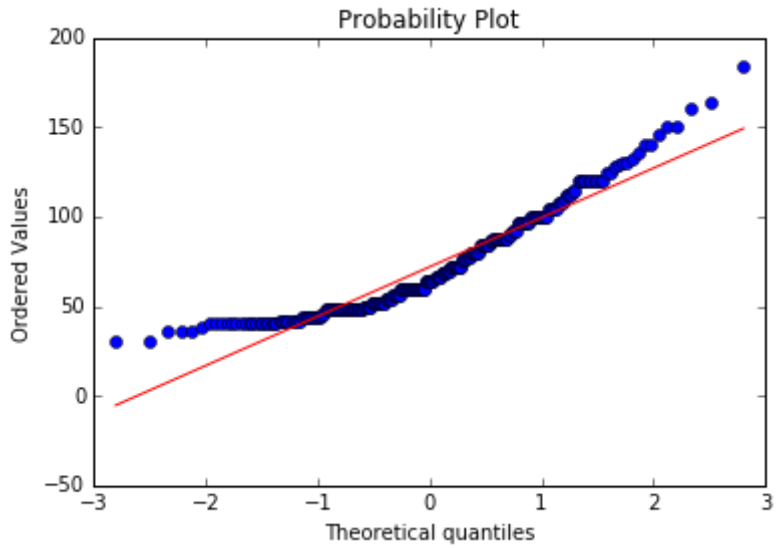
```
rectal_temperature
```



In [230]:

```
#qq图
attrid=5
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=stats.probplot(series, dist="norm", plot=pylab)
```

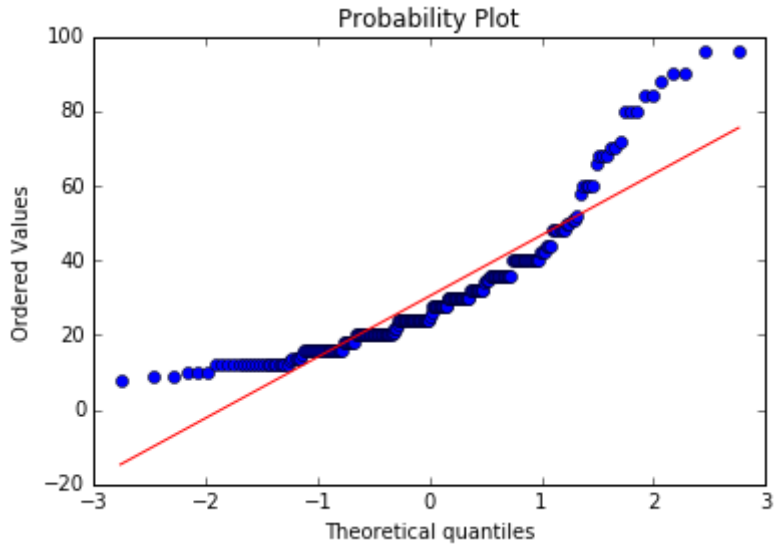
```
pulse
```



In [229]:

```
#qq图
attrid=6
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=stats.probplot(series, dist="norm", plot=pylab)
```

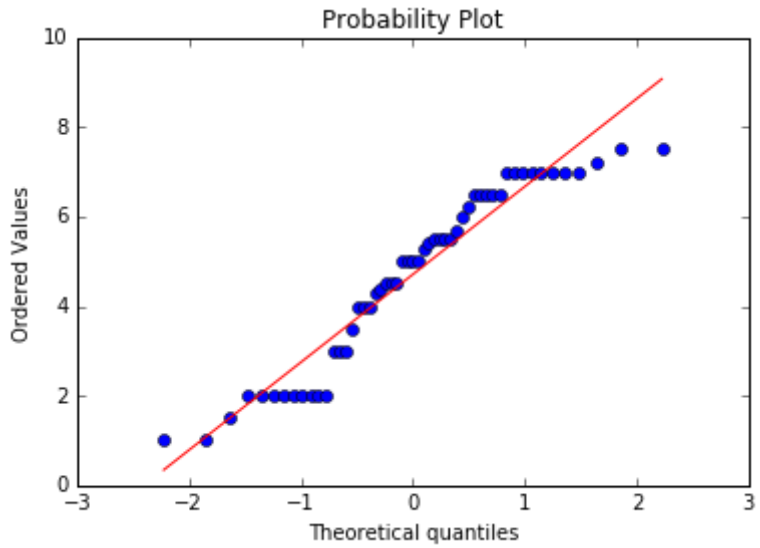
```
respiratory_rate
```



In [228]:

```
#qq图
attrid=16
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=stats.probplot(series, dist="norm", plot=pylab)
```

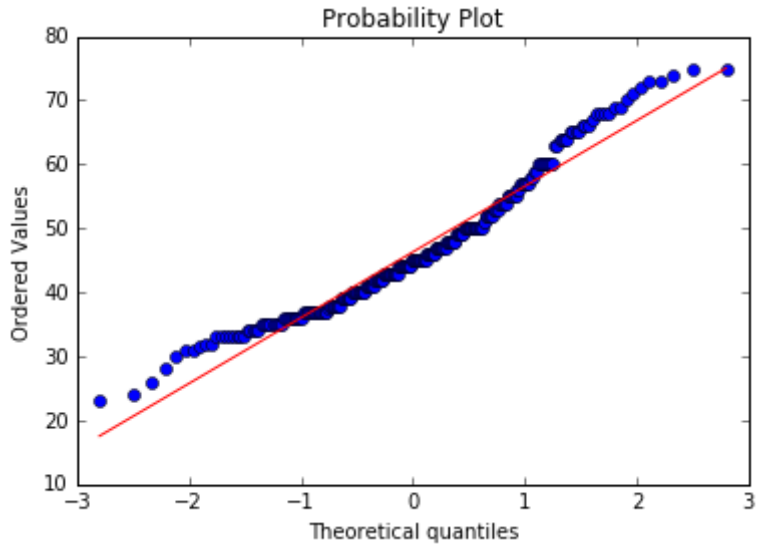
```
nasogastric_reflux_PH
```



In [227]:

```
#qq图
attrid=19
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=stats.probplot(series, dist="norm", plot=pylab)
```

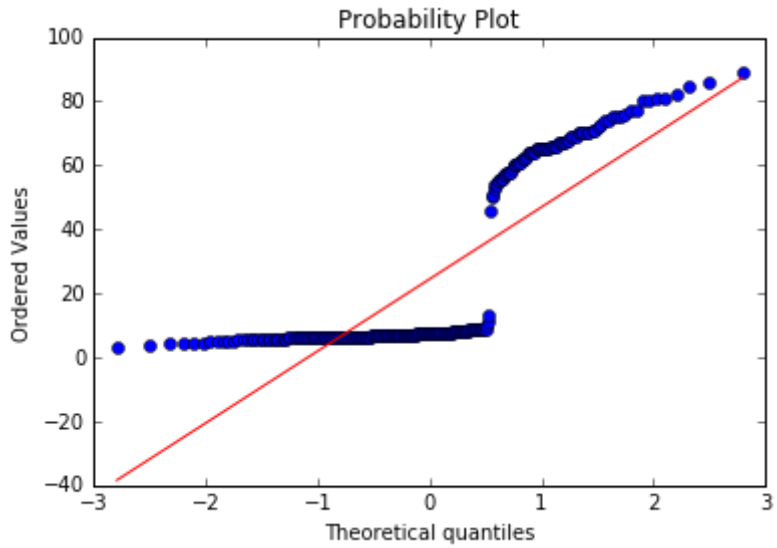
```
packed_cell_volume
```



In [226]:

```
#qq图
attrid=20
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=stats.probplot(series, dist="norm", plot=pylab)
```

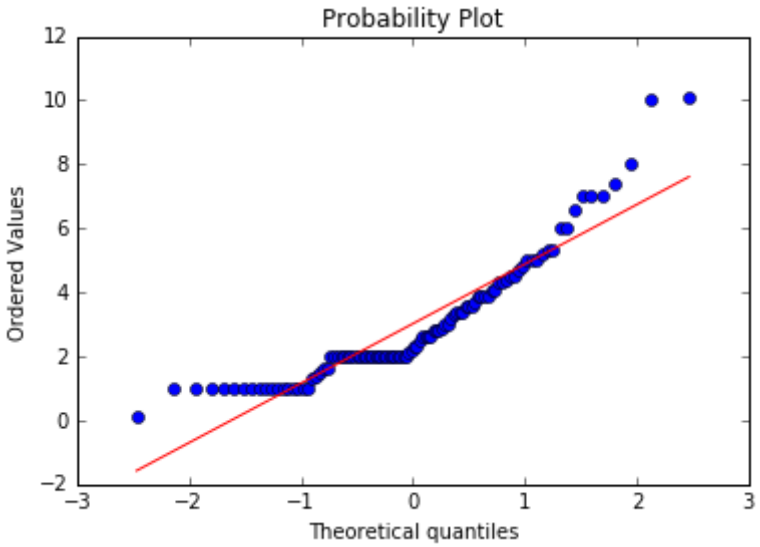
```
total_protein
```



In [233]:

```
#qq图
attrid=22
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=stats.probplot(series, dist="norm", plot=pylab)
```

```
abdomcentesis_total_protein
```

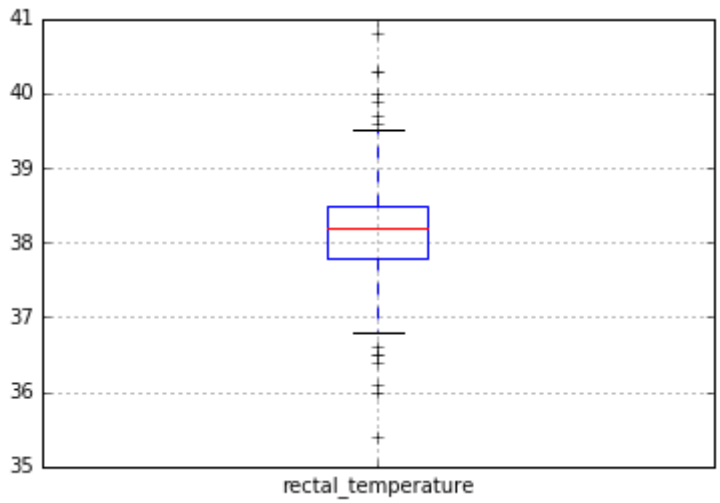


In [264]:

```
#盒图
attrid=4
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=pd.DataFrame(series).boxplot()
```

```
rectal_temperature
```

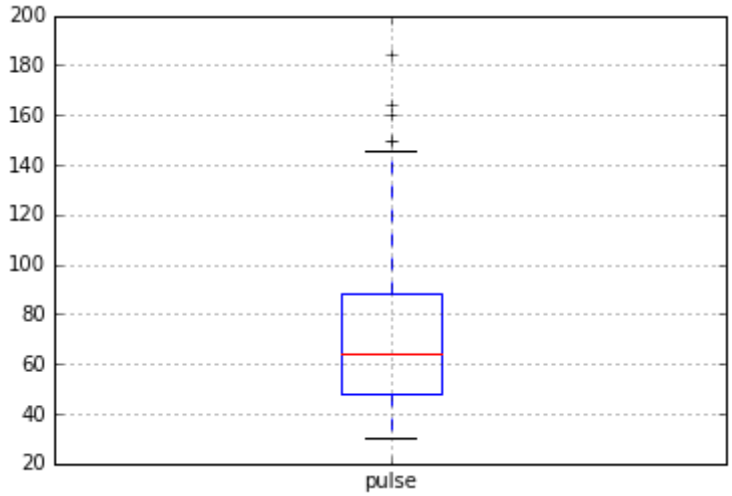
```
/home/mmjiang/.anaconda/lib/python3.5/site-packages/ipykernel/__main_
_.py:5: FutureWarning:
The default value for 'return_type' will change to 'axes' in a future
 release.
 To use the future behavior now, set return_type='axes'.
 To keep the previous behavior and silence this warning, set return_ty
pe='dict'.
```



In [265]:

```
#qq图
attrid=5
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=pd.DataFrame(series).boxplot()
```

```
pulse
/home/mmjiang/.anaconda/lib/python3.5/site-packages/ipykernel/__main_
_.py:5: FutureWarning:
The default value for 'return_type' will change to 'axes' in a future
 release.
 To use the future behavior now, set return_type='axes'.
 To keep the previous behavior and silence this warning, set return_ty
pe='dict'.
```

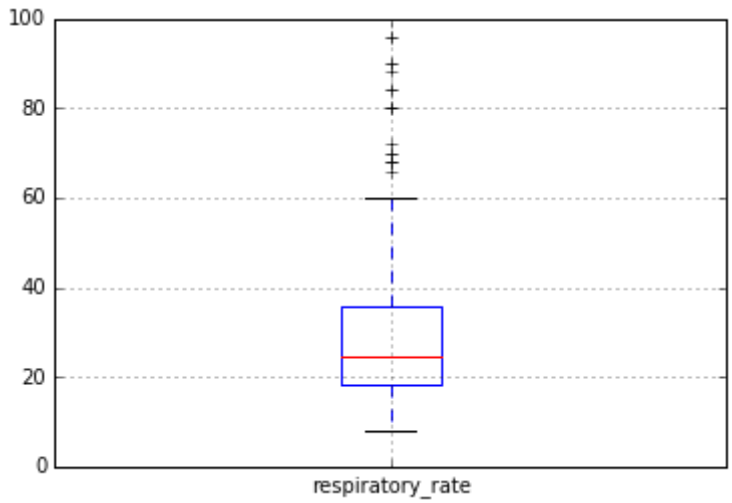


In [266]:

```
#qq图
attrid=6
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=pd.DataFrame(series).boxplot()
```

```
respiratory_rate
```

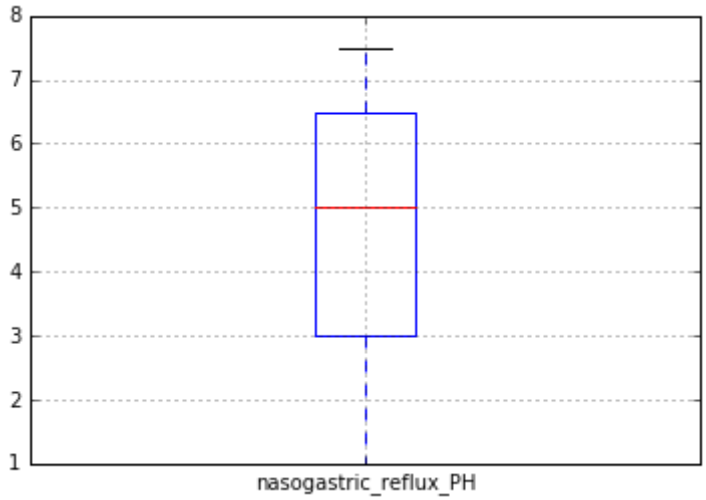
```
/home/mmjiang/.anaconda/lib/python3.5/site-packages/ipykernel/__main_
_.py:5: FutureWarning:
The default value for 'return_type' will change to 'axes' in a future
 release.
 To use the future behavior now, set return_type='axes'.
 To keep the previous behavior and silence this warning, set return_ty
pe='dict'.
```



In [267]:

```
#qq图
attrid=16
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=pd.DataFrame(series).boxplot()
```

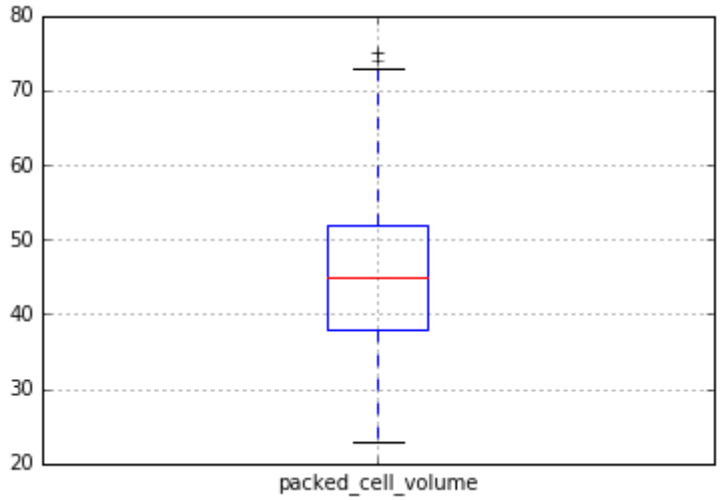
```
nasogastric_reflux_PH

/home/mmjiang/.anaconda/lib/python3.5/site-packages/ipykernel/__main_
_.py:5: FutureWarning:
The default value for 'return_type' will change to 'axes' in a future
 release.
 To use the future behavior now, set return_type='axes'.
 To keep the previous behavior and silence this warning, set return_ty
pe='dict'.
```



In [268]:

```
#qq图
attrid=19
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=pd.DataFrame(series).boxplot()
```

```
packed_cell_volume
/home/mmjiang/.anaconda/lib/python3.5/site-packages/ipykernel/__main_
_.py:5: FutureWarning:
The default value for 'return_type' will change to 'axes' in a future
 release.
 To use the future behavior now, set return_type='axes'.
 To keep the previous behavior and silence this warning, set return_ty
pe='dict'.
```



In [269]:

```
#qq图
attrid=20
print(attrname[attrid - 1])
series=df[attrname[attrid - 1]]
series=series[series != '?'].apply(pd.to_numeric, errors='coerce')
_=pd.DataFrame(series).boxplot()
```

```
total_protein
```

```
/home/mmjiang/.anaconda/lib/python3.5/site-packages/ipykernel/__main_
_.py:5: FutureWarning:
The default value for 'return_type' will change to 'axes' in a future
 release.
 To use the future behavior now, set return_type='axes'.
 To keep the previous behavior and silence this warning, set return_ty
pe='dict'.
```