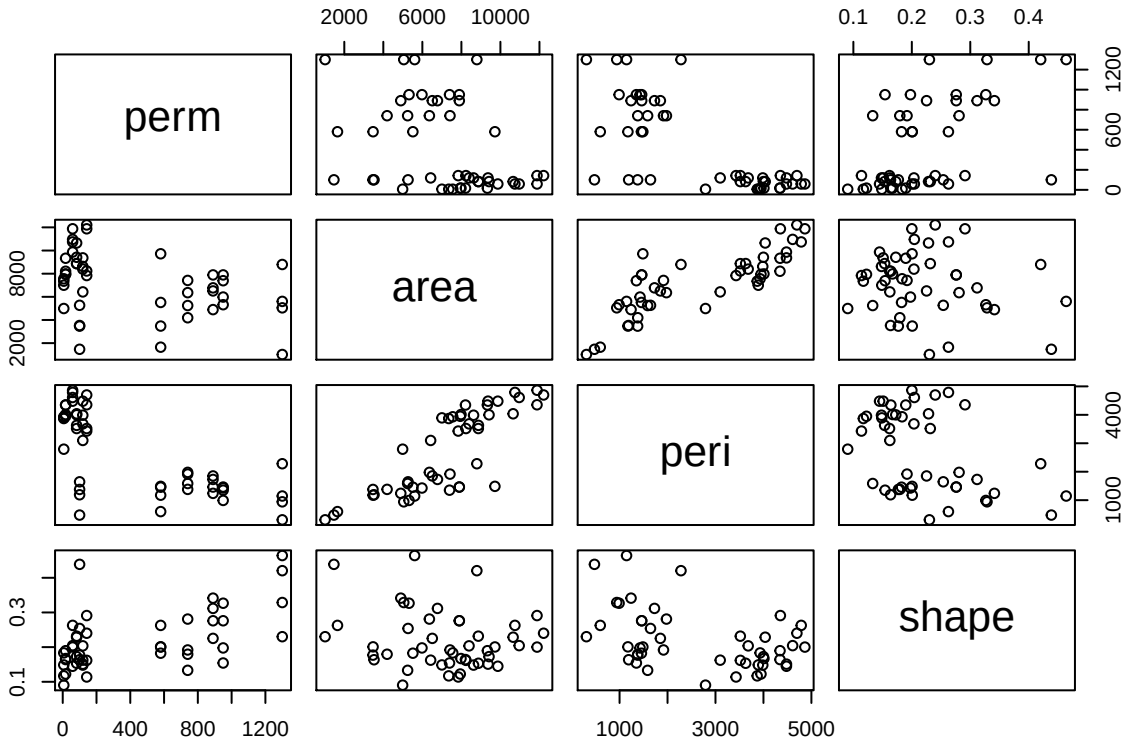
HW 2

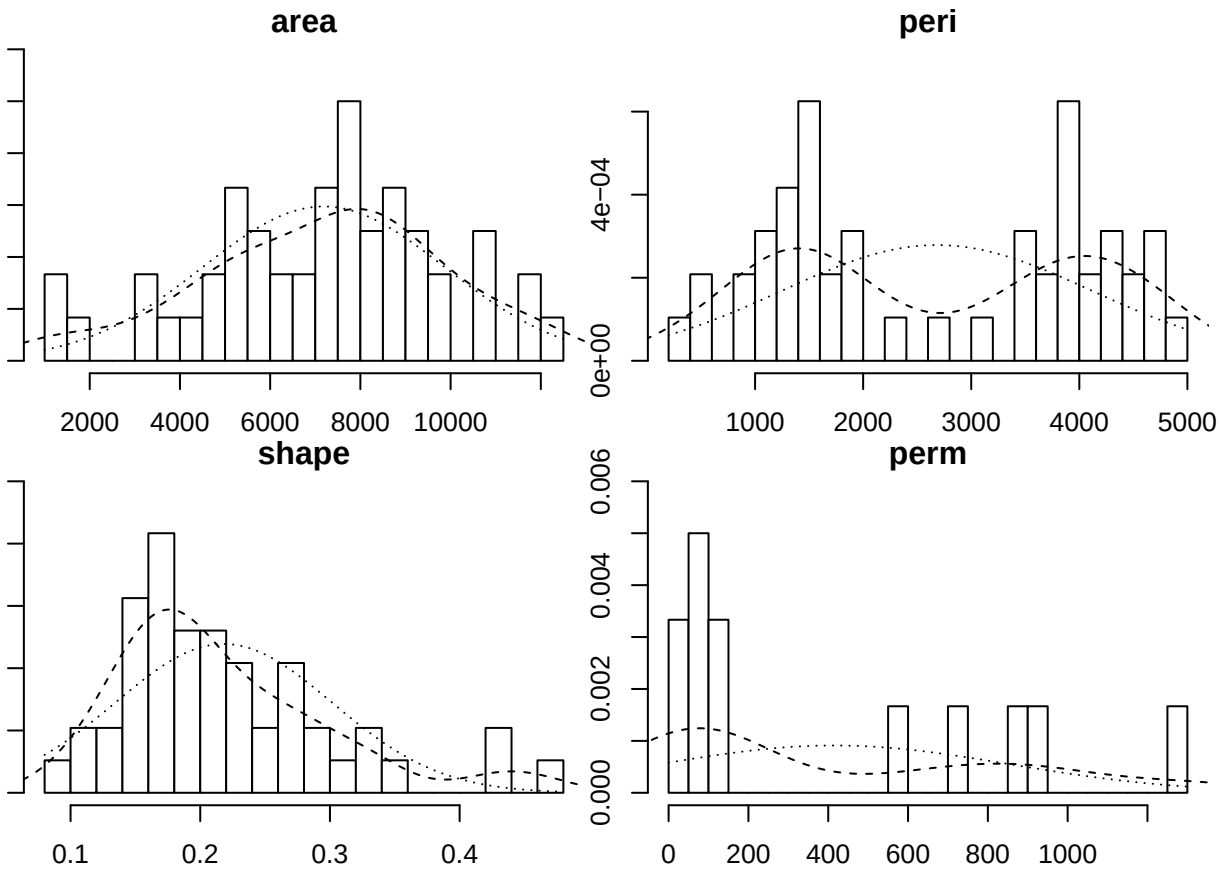
Ted Henson

1/27/2020

Question 2




```
##
## Attaching package: 'psych'
## The following object is masked from 'package:faraway':
##
##     logit
```




```
##
## Call:
## lm(formula = perm ~ ., data = rock)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -750.26  -59.57   10.66  100.25  620.91
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 485.61797  158.40826   3.066 0.003705 **
## area          0.09133    0.02499   3.654 0.000684 ***
## peri         -0.34402    0.05111  -6.731 2.84e-08 ***
## shape       899.06926  506.95098   1.773 0.083070 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 246 on 44 degrees of freedom
## Multiple R-squared:  0.7044, Adjusted R-squared:  0.6843
## F-statistic: 34.95 on 3 and 44 DF,  p-value: 1.033e-11
```

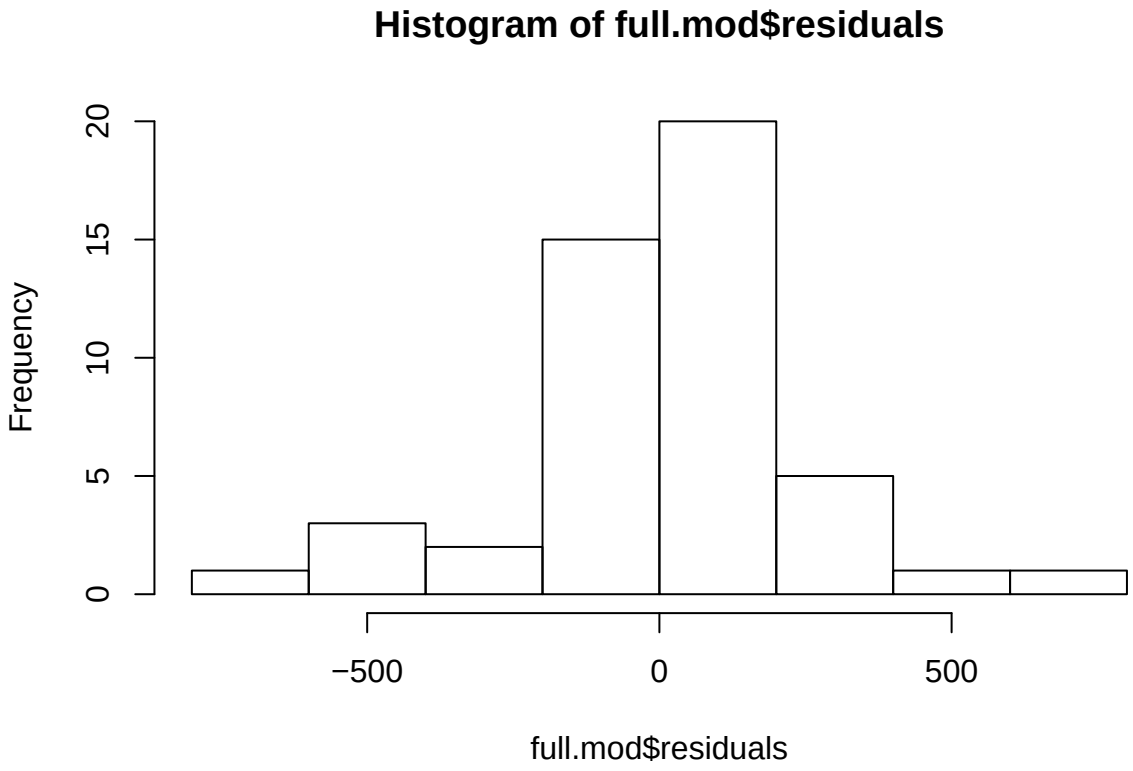
```
##                     2.5 %       97.5 %
## (Intercept)  166.36710209  804.8688468
## area           0.04096171    0.1417059
## peri          -0.44703814   -0.2410111
## shape       -122.62330057 1920.7618203
```

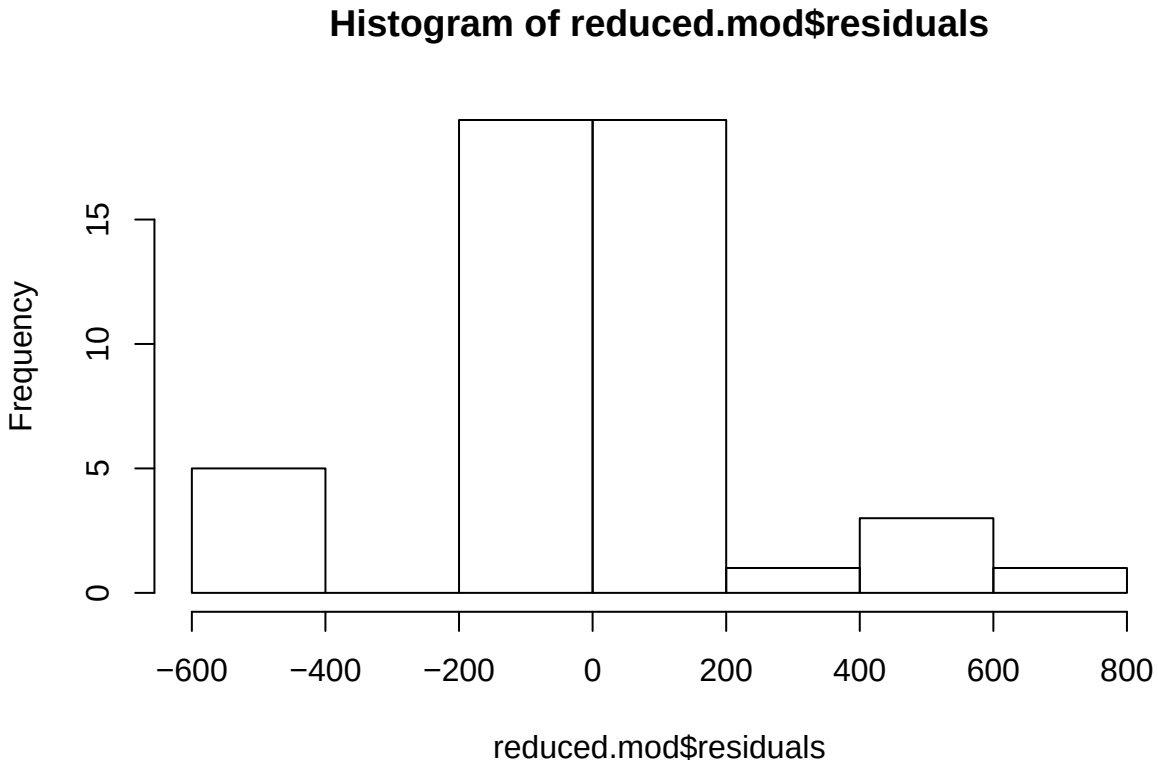
Stepwise AIC Model

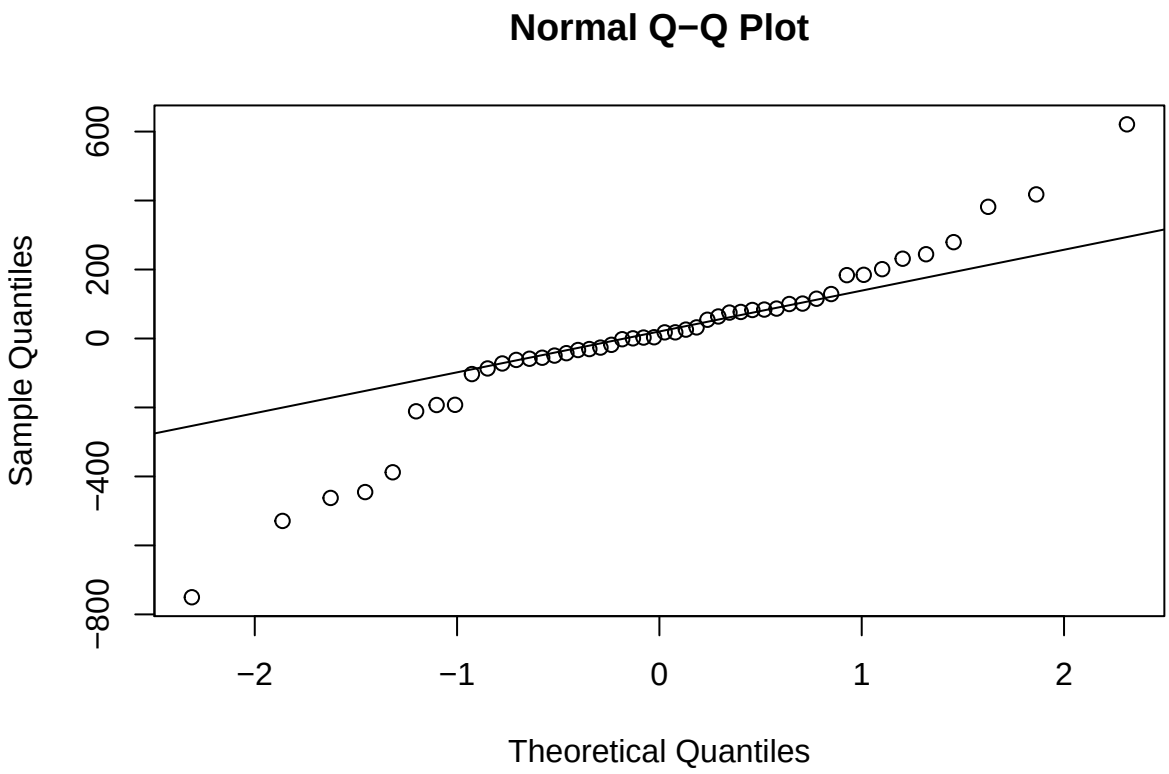
```
## Start:  AIC=584.84
## perm ~ 1
##
##         Df Sum of Sq     RSS    AIC
## + peri   1   4916322 4092864 548.97
## + shape  1   2792290 6216896 569.04
## + area   1   1417333 7591852 578.63
## <none>               9009186 584.84
##
## Step:  AIC=548.97
## perm ~ peri
##
##         Df Sum of Sq     RSS    AIC
## + area   1   1239481 2853383 533.66
## + shape  1    621651 3471213 543.06
## <none>               4092864 548.97
## - peri   1   4916322 9009186 584.84
##
## Step:  AIC=533.66
## perm ~ peri + area
##
##         Df Sum of Sq     RSS    AIC
## + shape  1    190360 2663023 532.34
## <none>               2853383 533.66
## - area   1   1239481 4092864 548.97
## - peri   1   4738469 7591852 578.63
##
## Step:  AIC=532.34
## perm ~ peri + area + shape
##
##         Df Sum of Sq     RSS    AIC
## <none>               2663023 532.34
## - shape  1    190360 2853383 533.66
## - area   1    808191 3471213 543.06
## - peri   1   2741707 5404730 564.32
##
## Call:
## lm(formula = perm ~ peri + area + shape, data = rock)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -750.26  -59.57   10.66  100.25  620.91
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 485.61797  158.40826   3.066 0.003705 **
## peri         -0.34402    0.05111  -6.731 2.84e-08 ***
## area          0.09133    0.02499   3.654 0.000684 ***
## shape       899.06926  506.95098   1.773 0.083070 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

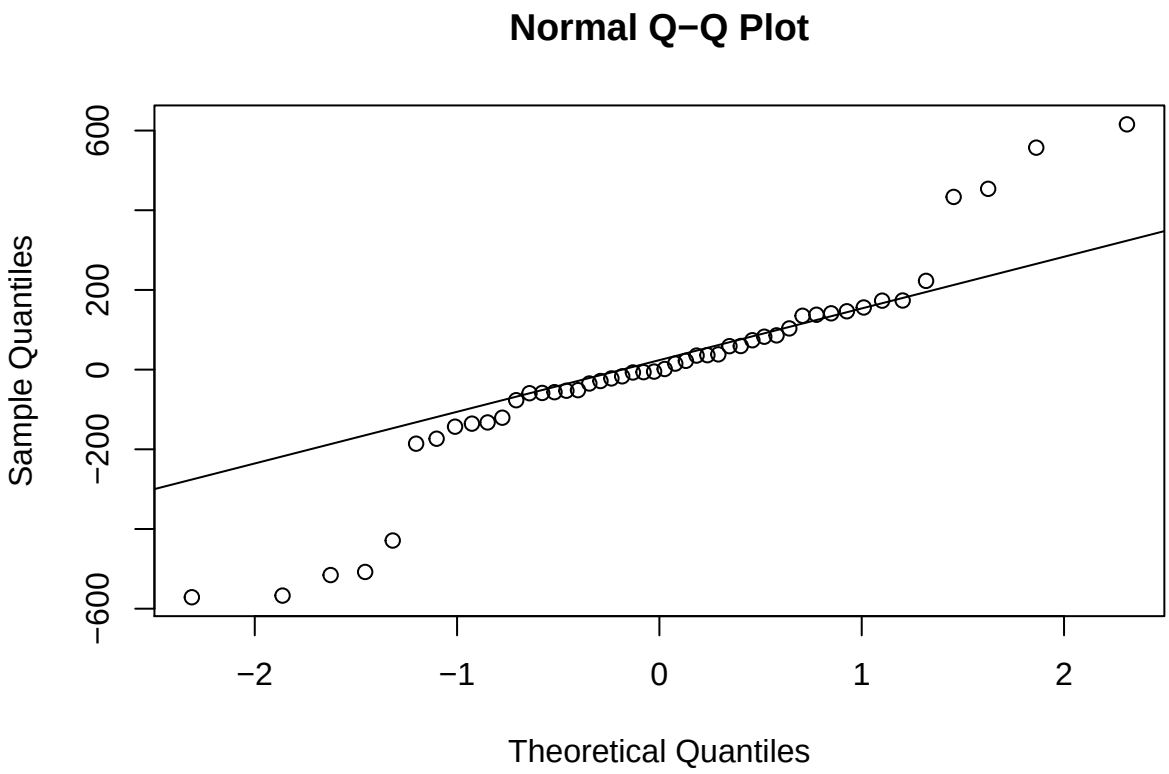
```
## Residual standard error: 246 on 44 degrees of freedom
## Multiple R-squared:  0.7044, Adjusted R-squared:  0.6843
## F-statistic: 34.95 on 3 and 44 DF,  p-value: 1.033e-11
```

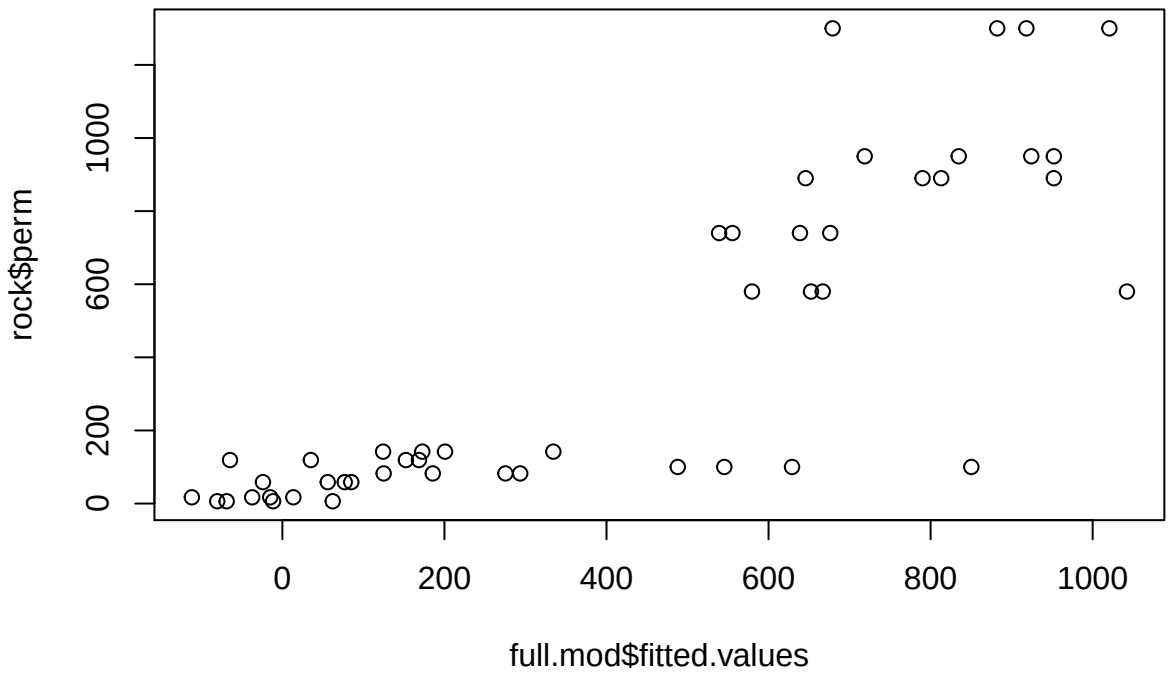
Model Diagnostics

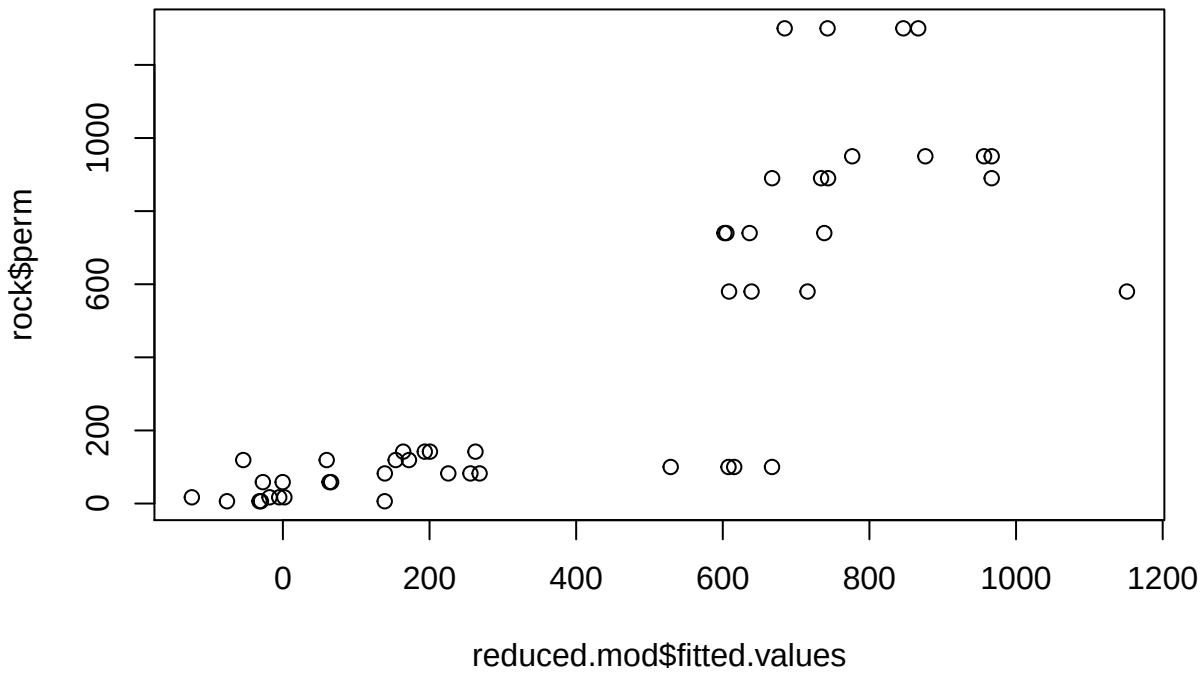





Normal Q−Q Plot
Sample Quantiles
600
200
0
−400
−800
−2
−1
0
1
2
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
600
200
0
−200
−600
−2
−1
0
1
2
Theoretical Quantiles

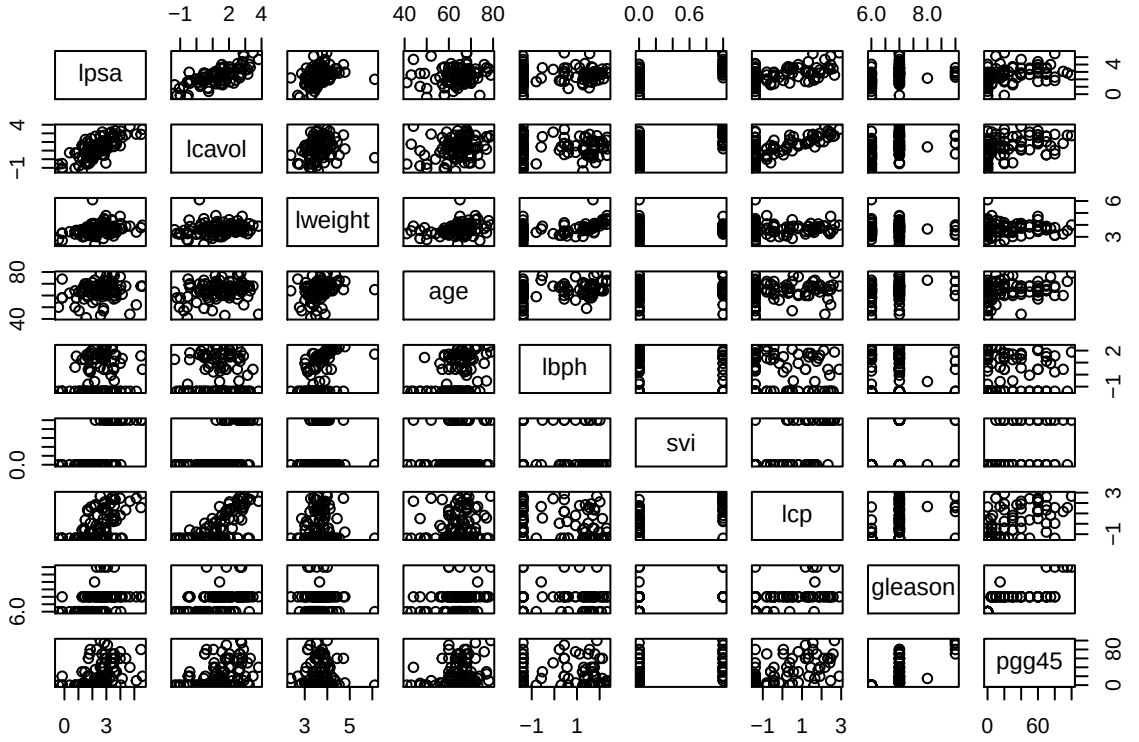

rock$perm
0
200
600
1000
0
200
400
600
800
1000
full.mod$fitted.values

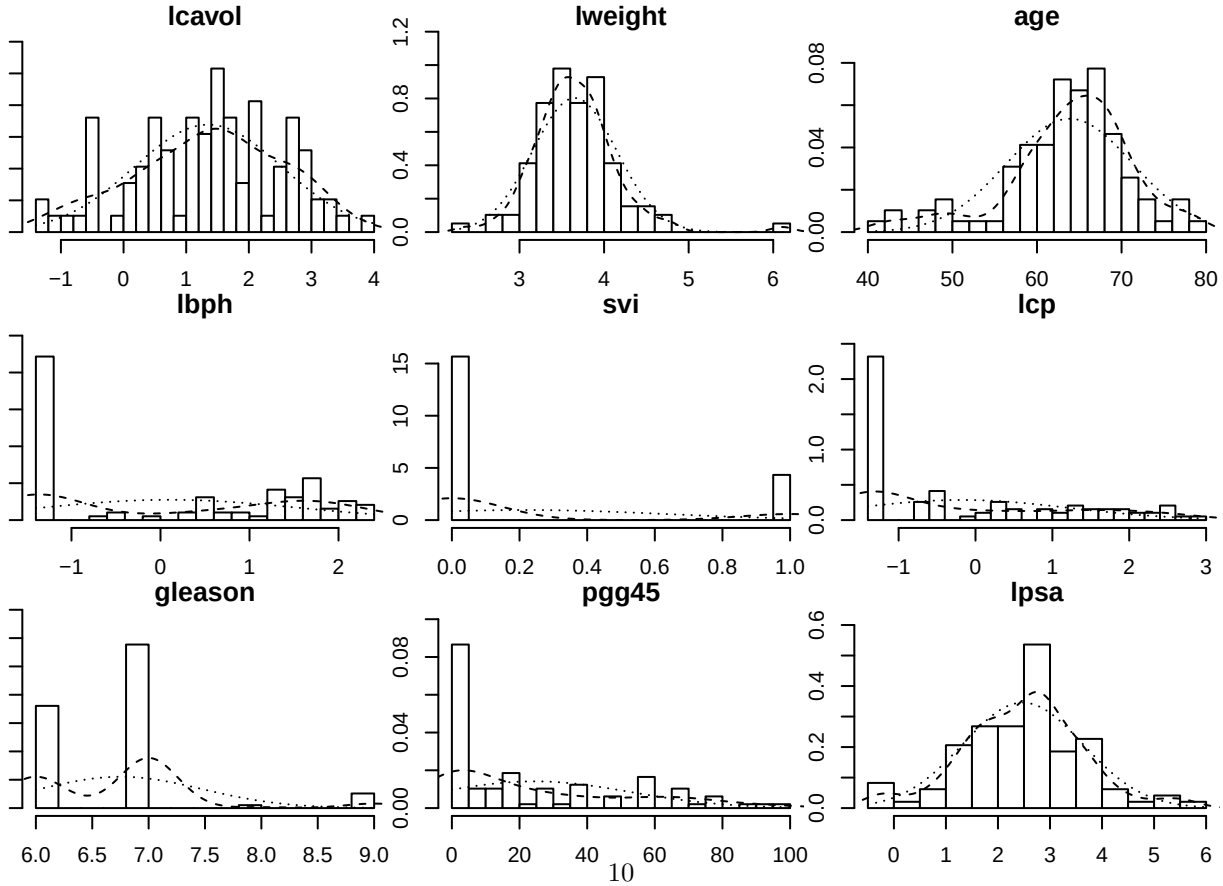


Question 2 Report

For the rock dataset, scatter plots of the predictor variables versus the response were constructed, as were histograms. The peri variable appeared to be bimodal and the response appeared to be skewed. The area and peri variables had highly significant p values and confidence intervals not including zero. The shape variable had a p value of about 8% and a wide confidence interval. The scatter plot was rather odd, but there still could be a relationship between shape and the response, perm. Stepwise regression did not eliminate this variable so a reduced model was built without the shape variable, and its residuals were compared to the full model. They were fairly similar, as were the plots of fitted versus actual values. It is difficult to say which model would perform better in practice, but one should probably go with the reduced model since it performed similarly with fewer variables.

Question 5


lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
lpsa

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
##                    2.5 %      97.5 %
## (Intercept) -1.906960983 3.245634379
## lcavol       0.412298699 0.761744954
## lweight      0.116603435 0.792331414
## age         -0.041840618 0.002566267
## lbph        -0.009101499 0.223209561
## svi          0.280644232 1.251670420
## lcp         -0.286344443 0.075395916
## gleason     -0.267786053 0.358069248
## pgg45       -0.004260932 0.013311395
```

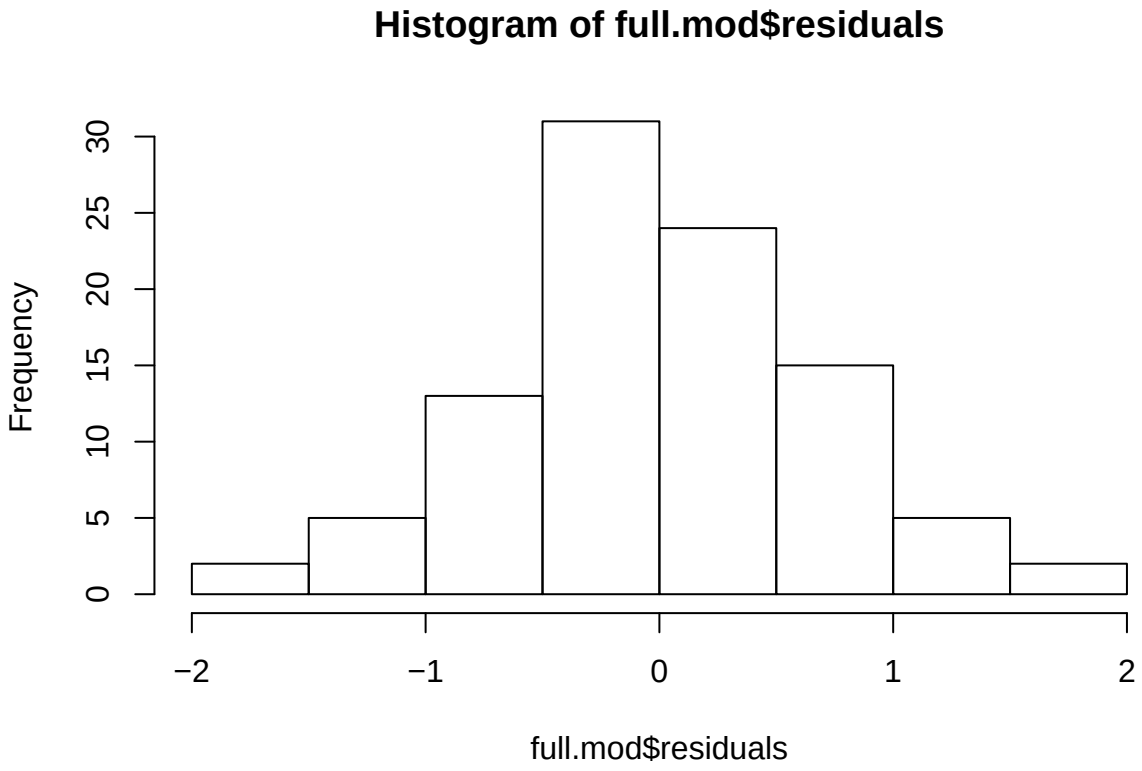
Stepwise AIC Model

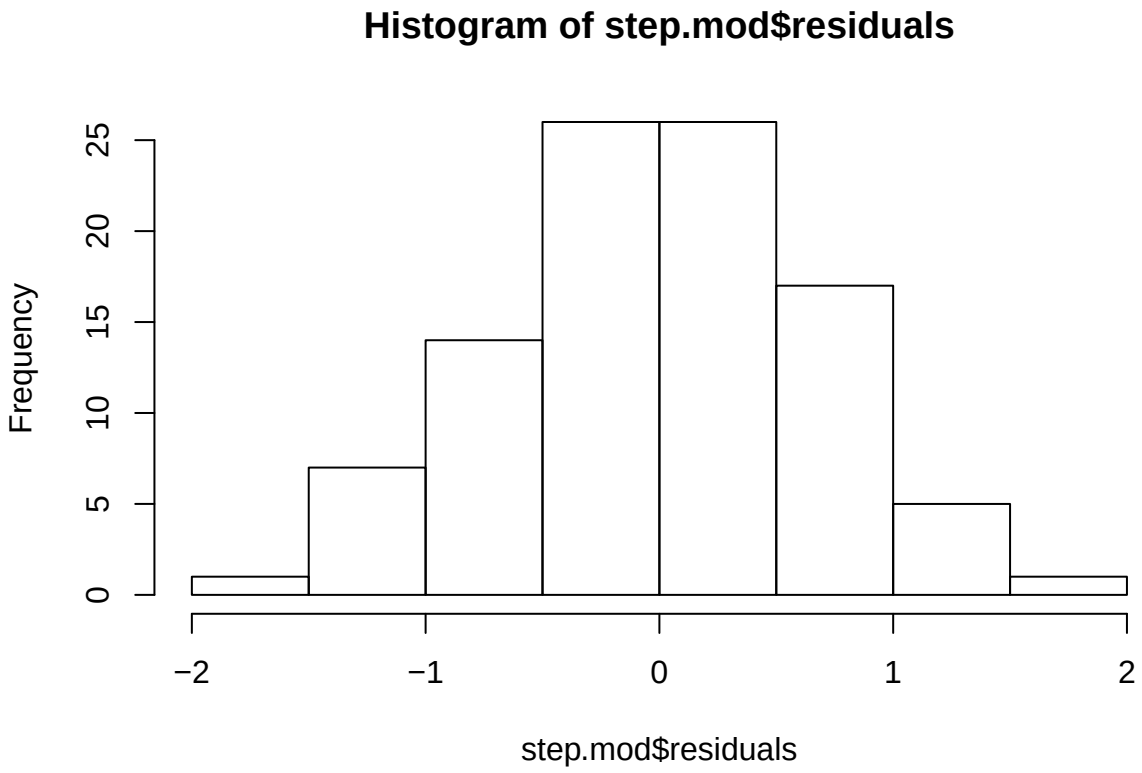
```
## Start:  AIC=-58.32
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  1    0.0412 44.204 -60.231
## - pgg45    1    0.5258 44.689 -59.174
## - lcp      1    0.6740 44.837 -58.853
## <none>                 44.163 -58.322
## - age      1    1.5503 45.713 -56.975
## - lbph     1    1.6835 45.847 -56.693
## - lweight  1    3.5861 47.749 -52.749
## - svi      1    4.9355 49.099 -50.046
## - lcavol   1   22.3721 66.535 -20.567
##
```

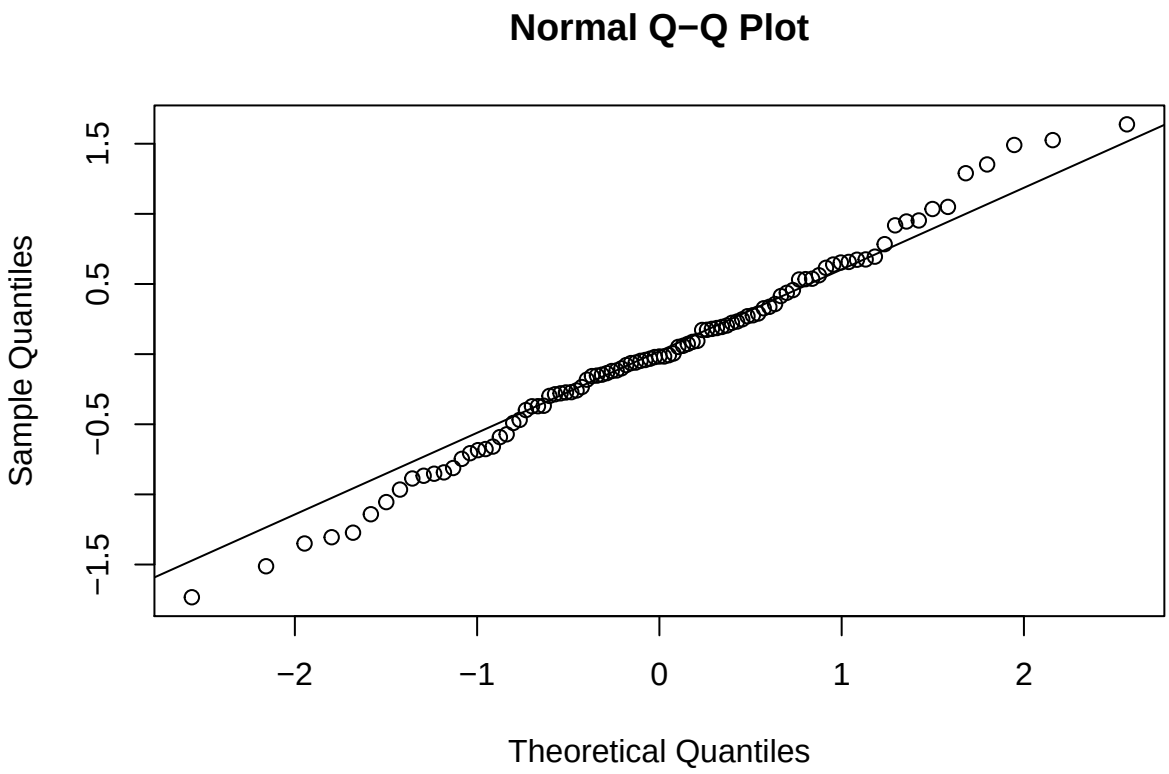
```
## Step:  AIC=-60.23
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.6623 44.867 -60.789
## <none>                  44.204 -60.231
## - pgg45    1    1.1920 45.396 -59.650
## - age      1    1.5166 45.721 -58.959
## - lbph     1    1.7053 45.910 -58.560
## + gleason  1    0.0412 44.163 -58.322
## - lweight  1    3.5462 47.750 -54.746
## - svi      1    4.8984 49.103 -52.037
## - lcavol   1   23.5039 67.708 -20.872
##
## Step:  AIC=-60.79
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.6590 45.526 -61.374
## <none>                  44.867 -60.789
## + lcp      1    0.6623 44.204 -60.231
## - age      1    1.2649 46.131 -60.092
## - lbph     1    1.6465 46.513 -59.293
## + gleason  1    0.0296 44.837 -58.853
## - lweight  1    3.5647 48.431 -55.373
## - svi      1    4.2503 49.117 -54.009
## - lcavol   1   25.4189 70.285 -19.248
##
## Step:  AIC=-61.37
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                  45.526 -61.374
## - age      1    0.9592 46.485 -61.352
## + pgg45    1    0.6590 44.867 -60.789
## + gleason  1    0.4560 45.070 -60.351
## + lcp      1    0.1293 45.396 -59.650
## - lbph     1    1.8568 47.382 -59.497
## - lweight  1    3.2251 48.751 -56.735
## - svi      1    5.9517 51.477 -51.456
## - lcavol   1   28.7665 74.292 -15.871
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
```

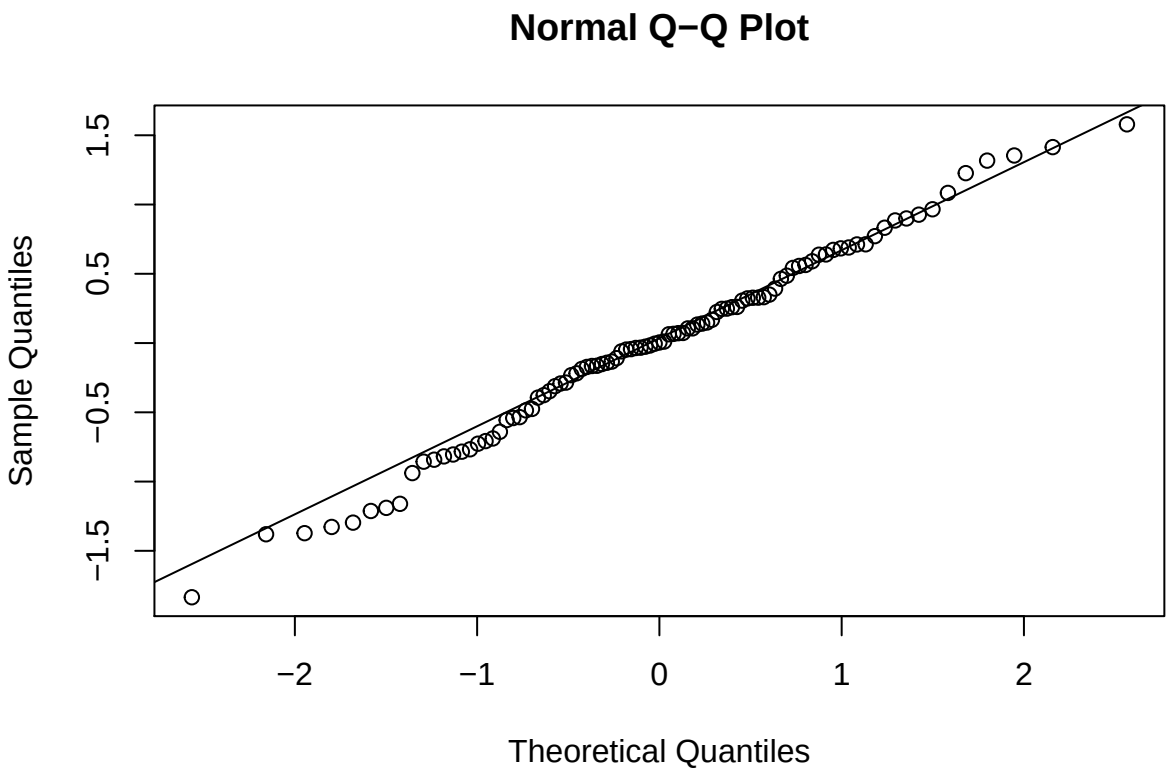
```
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

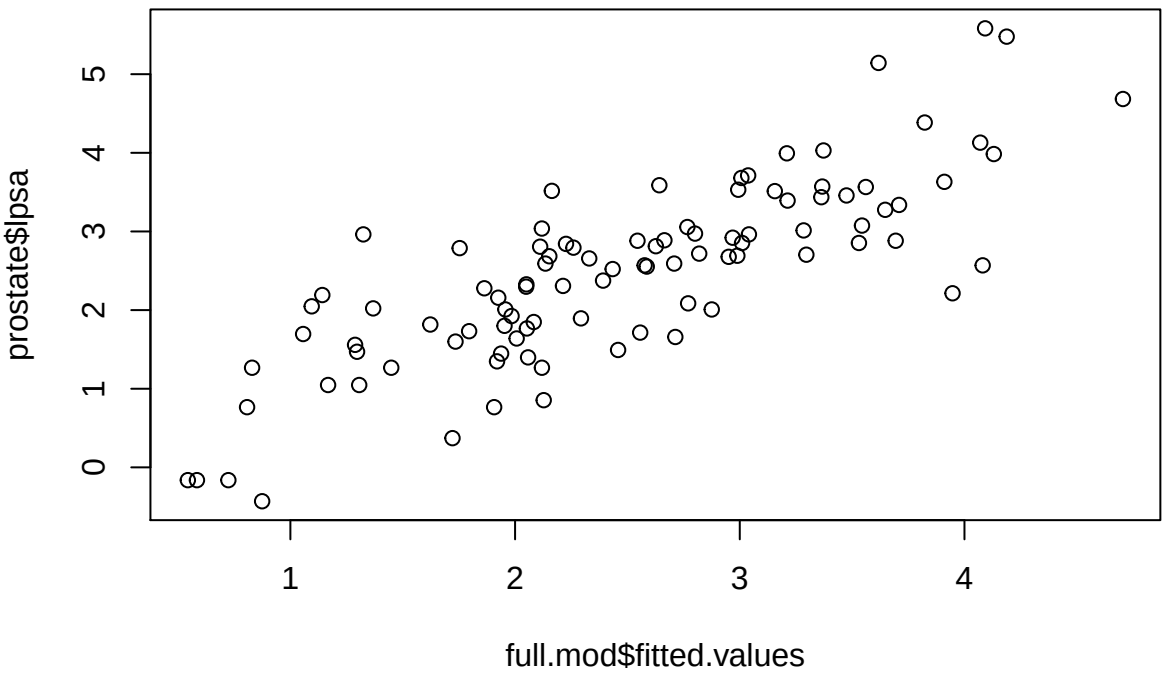
Model Diagnostics

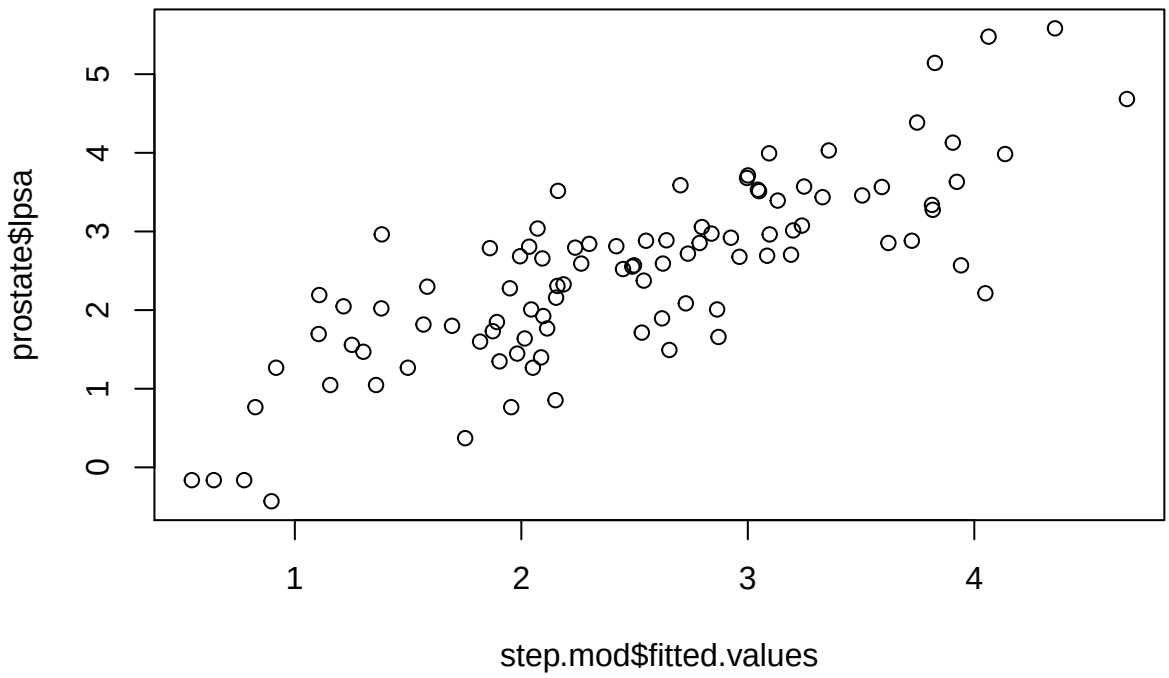





Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.5
0.5
−0.5
−1.5
−2
−1
0
1
2


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.5
0.5
−0.5
−1.5
−2
−1
0
1
2


prostate$lpsa
0
1
2
3
4
5
1
2
3
4
full.mod$fitted.values



Question 5 Report

For the prostate dataset, a similar process was constructed: scatter plots of the predictor variables and the response, along with histograms. In this case, the stepwise regression did eliminate some variables, mostly those with higher p values. In terms of the residuals and plots of fitted versus actual values, they were almost identical, so the reduced model should be chosen as it performed similarly with less predictor variables.