Homework 2

SOLUTIONS

Part A

1.

```
College.HS <- lm(College~HighSchool, data=state.data); College.HS
```

```
##
## Call:
## lm(formula = College ~ HighSchool, data = state.data)
##
## Coefficients:
## (Intercept)   HighSchool
##     -96.366        1.426
```

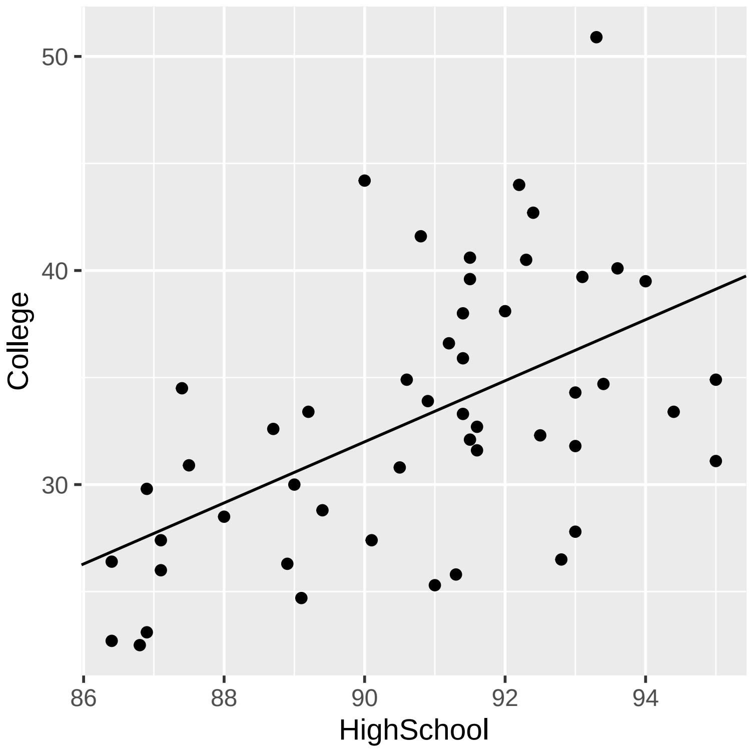
2.

For each additional percent of a state's population that earns a high school diploma, we expect college graduation percent to increase by 1.4%.

3.

Certainly a line seems appropriate here. There are some outliers (points far from the line) on the right side of the graph.

```
#Here you should have the code to make the scatterplot with line
gf_point(College~HighSchool, data=state.data)+ geom_abline(intercept=College.HS$coefficients[1], slope=C
```



4.

The output below shows an R^2 of 29.2%, which is the percent of the variation in college graduation percentage explained by this model.

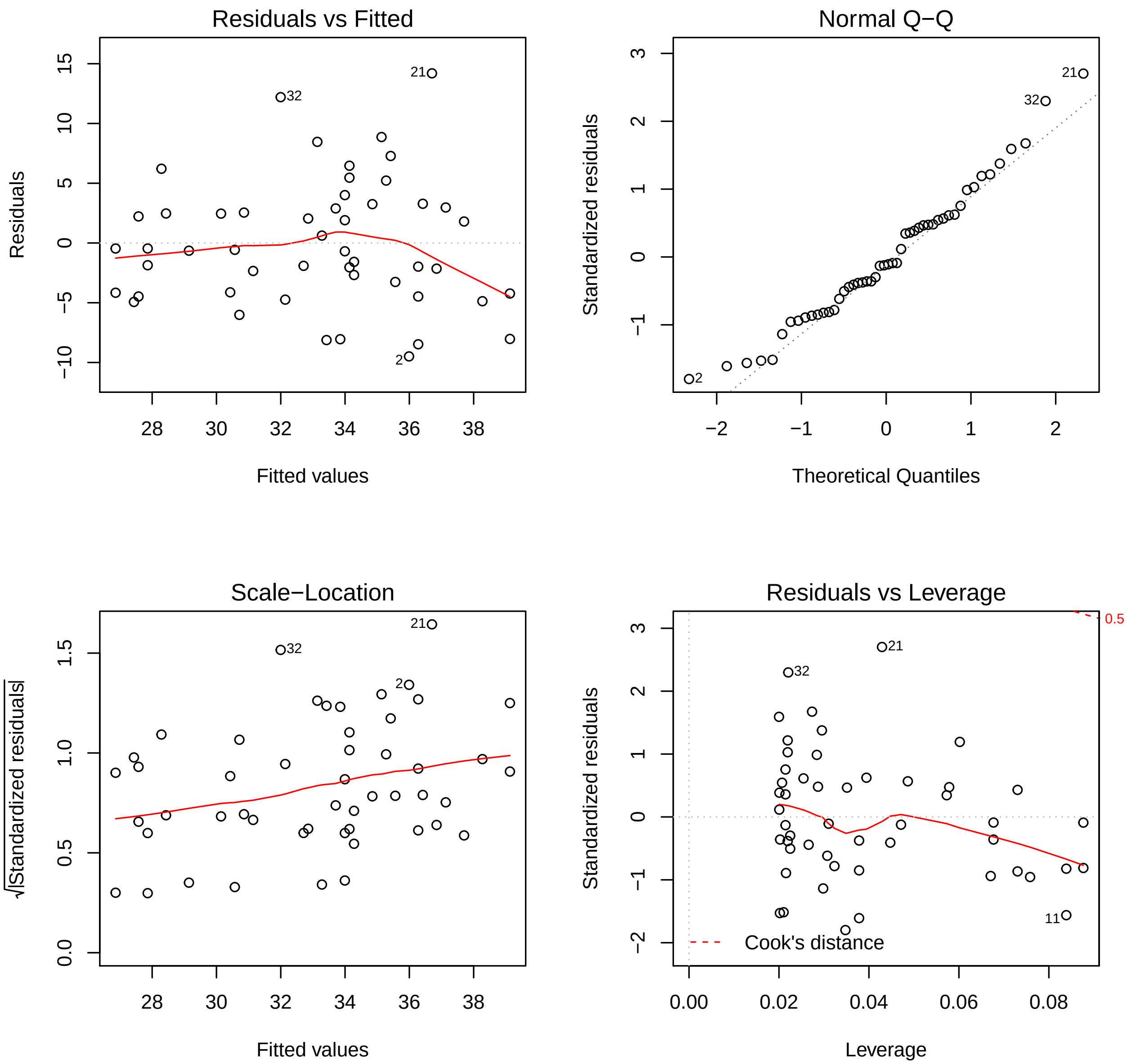
```
summary(College.HS)
```

```
##
## Call:
## lm(formula = College ~ HighSchool, data = state.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4907 -4.1541 -0.6078  2.9490 14.1962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -96.3657    29.0740  -3.314  0.00175 **
## HighSchool    1.4263     0.3202   4.454 5.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.37 on 48 degrees of freedom
## Multiple R-squared:  0.2924, Adjusted R-squared:  0.2777
## F-statistic: 19.84 on 1 and 48 DF,  p-value: 5.03e-05
```

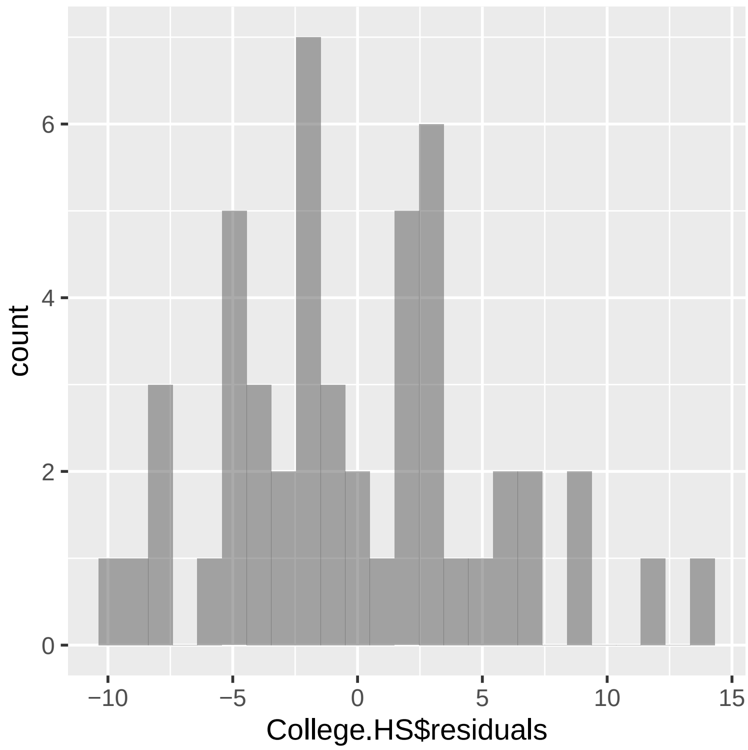
5.

Constant variance might be a concern, as shown in the first residual plot – it's hard to say becaause there isn't much data (only 50 points). Normality is fine – even though the histogram doesn't look good, the normal probability plot shows that the residuals are reasonably normal.

```
par(mfrow=c(2,2))
plot(College.HS)
```




```
gf_histogram(~College.HS$residuals)
```



Part B

1.

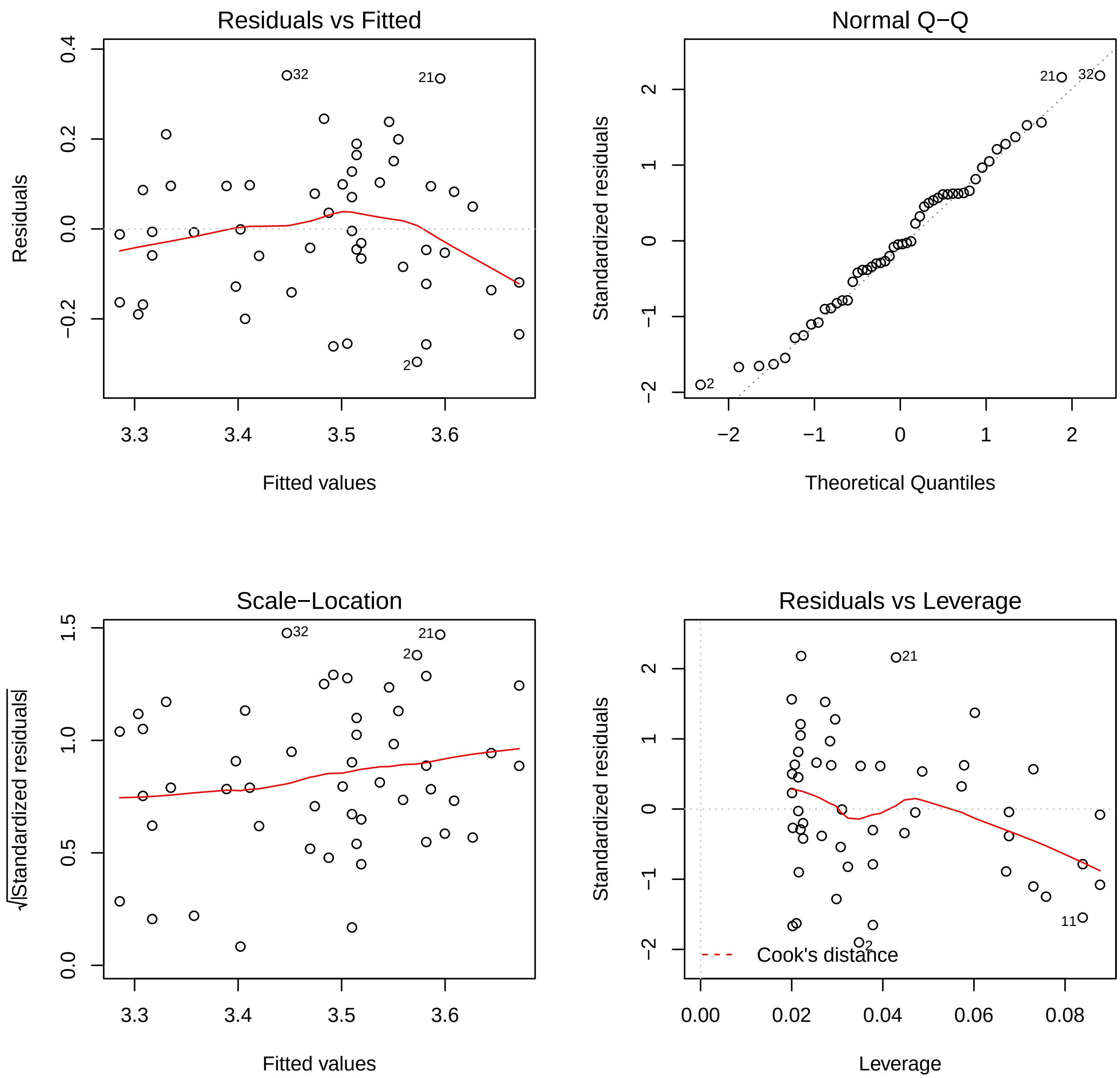
```
logCollege.HS <- lm(log(College)~HighSchool, data=state.data); logCollege.HS
```

```
##
## Call:
## lm(formula = log(College) ~ HighSchool, data = state.data)
##
## Coefficients:
## (Intercept)   HighSchool
##    -0.59195      0.04488
```

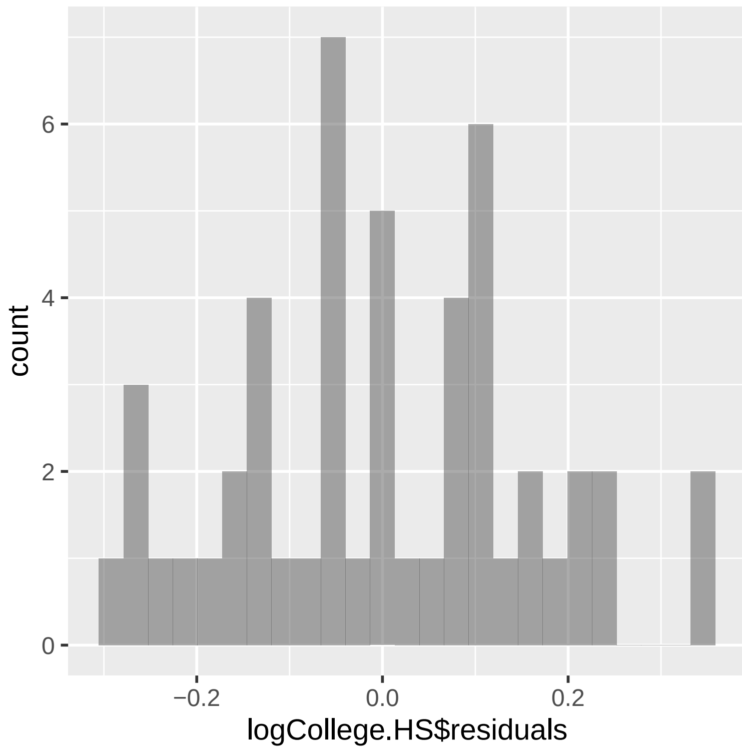
2.

This looks much better! Constant variance and normality both look to be met, and no influential points.

```
par(mfrow=c(2,2))
plot(logCollege.HS)
```




```
gf_histogram(~logCollege.HS$residuals)
```



3.

The test of slope=0 has a p-value of approx. 0, so yes, there is a significant association between log(College graduation %) and High School graduation % for US states.

```r
summary(logCollege.HS)
```

```
##
## Call:
## lm(formula = log(College) ~ HighSchool, data = state.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29568 -0.12152 -0.00699  0.09692  0.34156
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.591945   0.857195  -0.691    0.493
## HighSchool   0.044879   0.009441   4.754 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1583 on 48 degrees of freedom
## Multiple R-squared:  0.3201, Adjusted R-squared:  0.3059
## F-statistic:  22.6 on 1 and 48 DF,  p-value: 1.856e-05
```

4.

We are 90% confident that high school graduation rate increases by 1%, we expect log(College graduation %) to increaase between 0.029 and 0.061.

```
confint(logCollege.HS, level=0.9)
```

```
##                     5 %       95 %
## (Intercept) -2.02965407 0.84576382
## HighSchool   0.02904399 0.06071403
```

5.

The F-test (p-value=0.000019) tells us that High School graduation % is useful in predicting log(College graduation %).

```
anova(logCollege.HS)
```

```
## Analysis of Variance Table
##
## Response: log(College)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## HighSchool  1 0.56637 0.56637  22.596 1.856e-05 ***
## Residuals  48 1.20312 0.02506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part C

1.

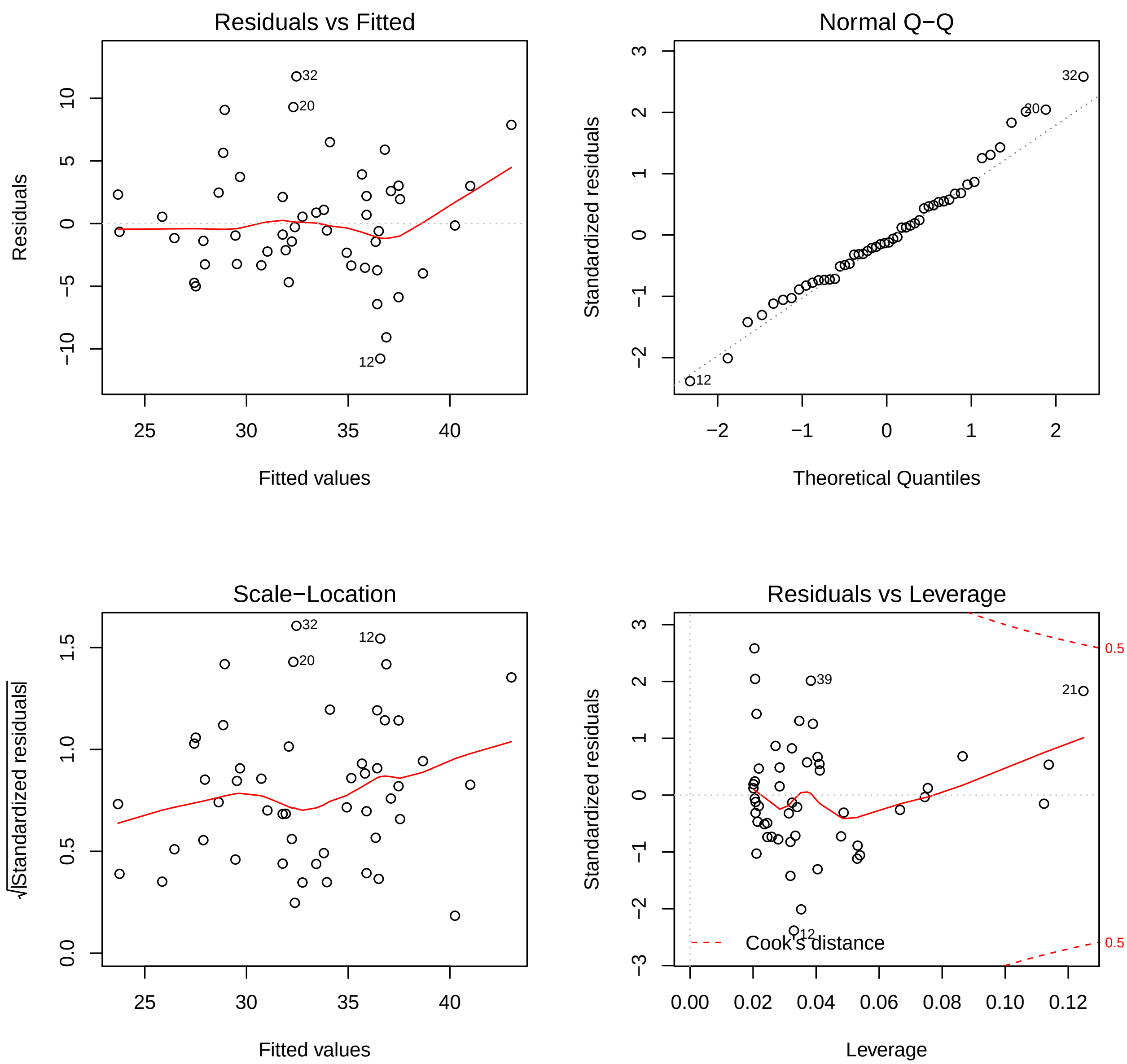
Below are the 3 models with the highest R^2 values. Of these, I think the percentage of vaccinated residents is the best predictor. It has the highest R^2 and no major problems with conditions. (If you ignore the outlier of point 31, you'll see that there is no issue with constant variance. And point 31 is not influential.)

```
par(mfrow=c(2,2))
College.8math <- lm(College~EighthGradeMath, data=state.data)
summary(College.8math)
```

```
##
## Call:
## lm(formula = College ~ EighthGradeMath, data = state.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7760  -3.2482  -0.5766   2.4326  11.7472
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -177.7598    31.5700  -5.631 9.14e-07 ***
## EighthGradeMath    0.7497     0.1122   6.680 2.28e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.596 on 48 degrees of freedom
## Multiple R-squared:  0.4818, Adjusted R-squared:  0.471
## F-statistic: 44.62 on 1 and 48 DF,  p-value: 2.277e-08
```

```
plot(College.8math)
```

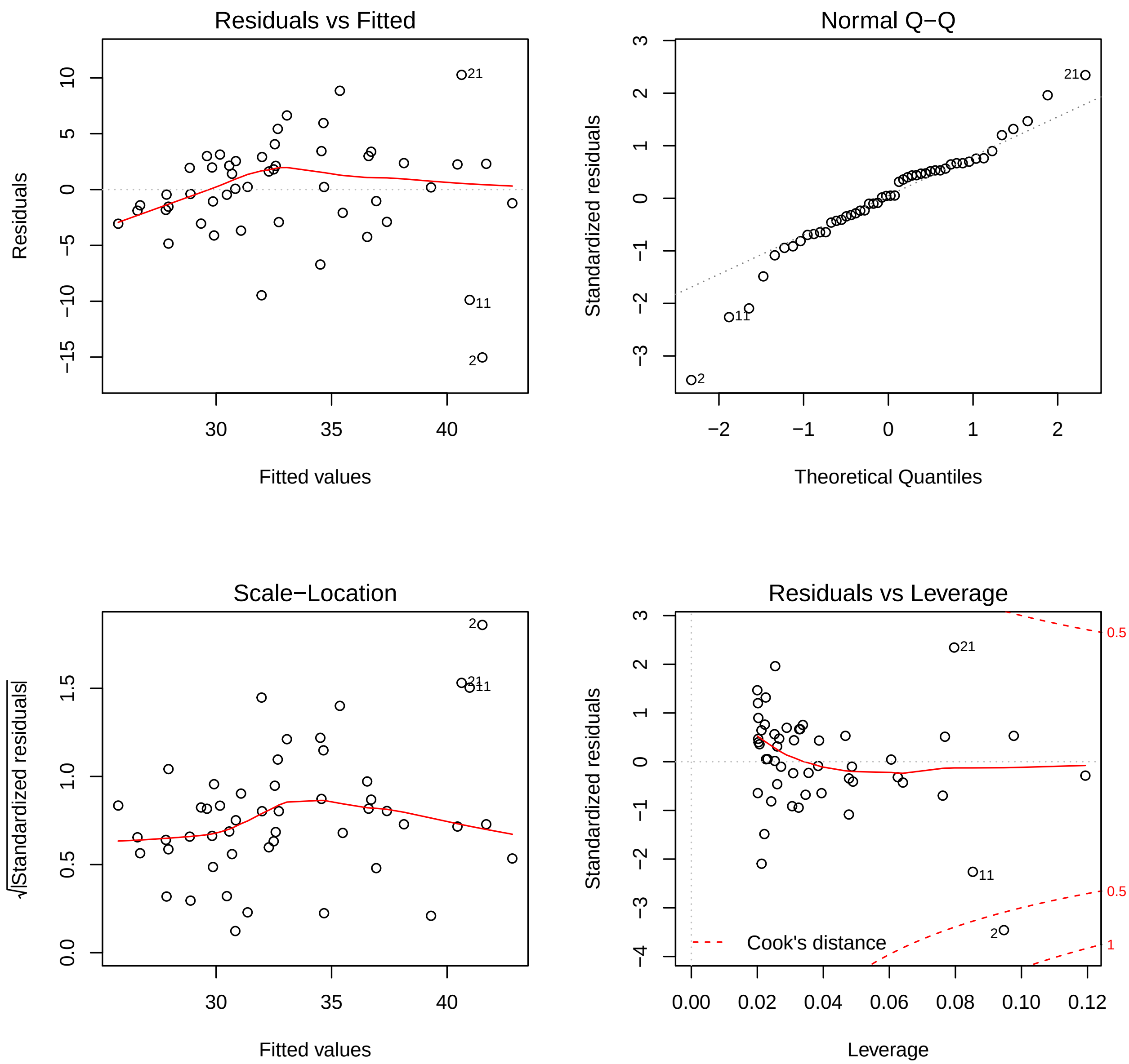



```
College.Income <- lm(College~HouseholdIncome, data=state.data)
summary(College.Income)
```

```
##
## Call:
## lm(formula = College ~ HouseholdIncome, data = state.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0295  -2.0356   0.2105   2.5019  10.2707
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.33689    4.00559   1.582     0.12
## HouseholdIncome   0.46237    0.06834   6.766 1.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.567 on 48 degrees of freedom
## Multiple R-squared:  0.4882, Adjusted R-squared:  0.4775
## F-statistic: 45.78 on 1 and 48 DF,  p-value: 1.68e-08
```

```r
plot(College.Income)
```

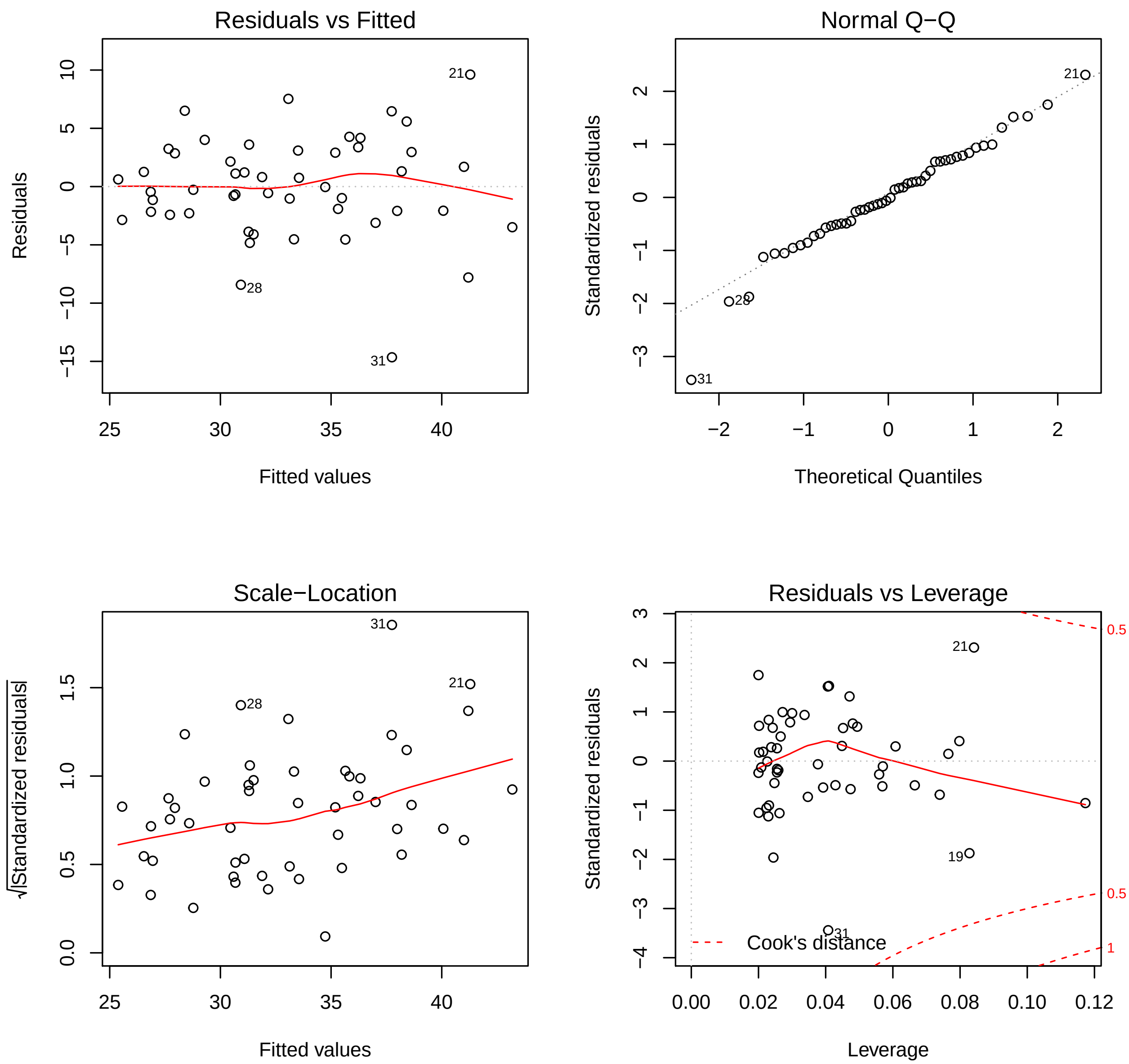



```r
College.vax <- lm(College~percent_fully_vax, data=state.data)
summary(College.vax)
```

```
##
```

```
## Call:
## lm(formula = College ~ percent_fully_vax, data = state.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6497  -2.2589  -0.1564   2.9509   9.6126
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.63663    3.60211   1.842   0.0716 .
## percent_fully_vax  0.53818    0.07222   7.452  1.5e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.347 on 48 degrees of freedom
## Multiple R-squared:  0.5363, Adjusted R-squared:  0.5267
## F-statistic: 55.52 on 1 and 48 DF,  p-value: 1.499e-09
```

```r
plot(College.vax)
```



2.

We are 95% confident that among states with a 48.27% vaccination rate, the mean college graduation percentage will be between 31.37% and 33.86%.

```
median(~percent_fully_vax, data=state.data)
```

```
## [1] 48.26808
```

```
#Code for confidence interval:
predict.lm(College.vax, data.frame("percent_fully_vax"=48.27), interval="confidence")
```

```
##        fit      lwr      upr
## 1 32.61436 31.37187 33.85684
```

3.

We are 95% confident that for a state with a 48.27% vaccination rate, the college graduation percentage will be between 23.79% and 41.44%.

```r
#Code for prediction interval:
predict.lm(College.vax, data.frame("percent_fully_vax"=48.27), interval="prediction")
```

```
##        fit      lwr     upr
## 1 32.61436 23.78661 41.4421
```