Assign6_Choi_Regression Diagnostics

Serena Choi

Setting up

```
chirot <- read.csv("chirot.csv")
attach(chirot)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
library(car)
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
library(carData)
```

Regression model

Model 1 below is a linear regression model with rebellion intensity ("intensity") as the dependent variable and commercialization ("commercial"), traditionalism ("tradition"), strength of middle class peasants ("peasant"), and land inequality ("inequal") as the independent variables.

```
Model1 <- lm(intensity ~ commercial + tradition + peasant + inequal, data = chirot)
summary(Model1)
```

```
##
## Call:
## lm(formula = intensity ~ commercial + tradition + peasant + inequal,
##     data = chirot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2460 -0.6781 -0.1013  0.8025  2.3378
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.919018   5.507499  -2.346 0.026587 *
## commercial    0.091140   0.020268   4.497 0.000118 ***
## tradition     0.116787   0.060688   1.924 0.064906 .
## peasant      -0.003342   0.017695  -0.189 0.851625
## inequal       1.137970   2.850304   0.399 0.692853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 27 degrees of freedom
## Multiple R-squared:  0.5836, Adjusted R-squared:  0.5219
```

```
## F-statistic: 9.462 on 4 and 27 DF,  p-value: 6.476e-05
```

According to the Model 1, independent variables including commercialization, traditionalism and inequality are positively associated with rebellion intensity. On the other hand strength of middle class peasantry is negatively associated with rebellion intensity. However, commercialization is the only variable that is statistically significant to rebellion intensity. The fit of the model is good as the adjusted R-squared is quite high.

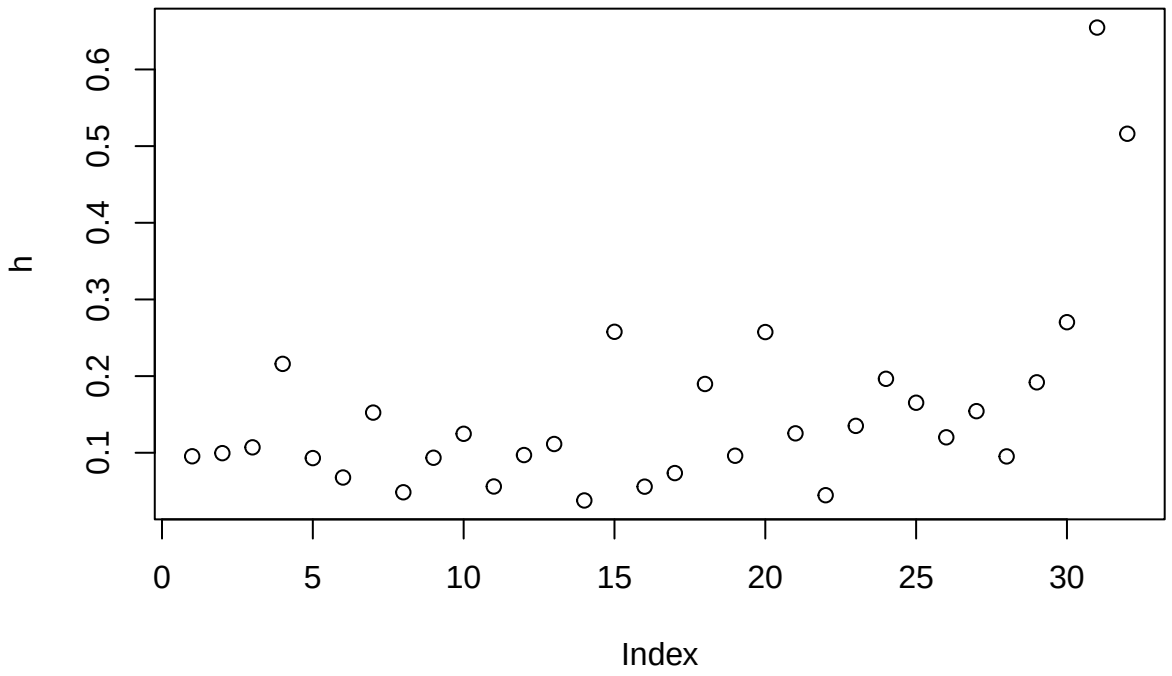
Outliers, leverage and influence

Assess the impact of any discrepant observations on your model estimates using the diagnostic tools for leverage, outliers, and influence we have discussed in class. Did you find any influential observations?

Detecting leverage

Leverage points are observations that are outlying in the x-direction. In order to detect them, we will examine the hat values, which capture contribution of each observation to the fitted values.

```
h<-hatvalues(Model1)
plot(h)
```

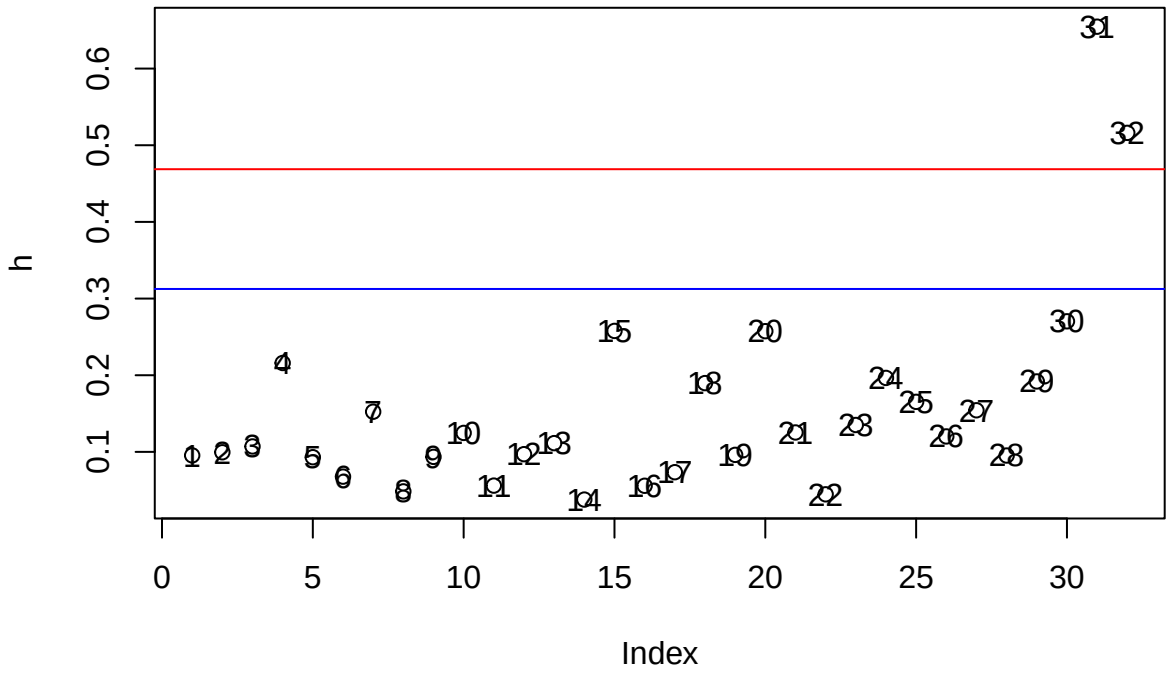



```
2*mean(h)
```

```
## [1] 0.3125
```

Looking at the scatter plot above, there are points greater than twice the average of hat values. Therefore, we can suspect that there are leverage points. We will further confirm this suspicion by calculating the hat values.

```
k<-4
n<-32
plot(h)
text(h, label=(1:length(county)))
abline(h=(2*(k+1)/n), col="blue")
abline(h=(3*(k+1)/n), col="red")
```

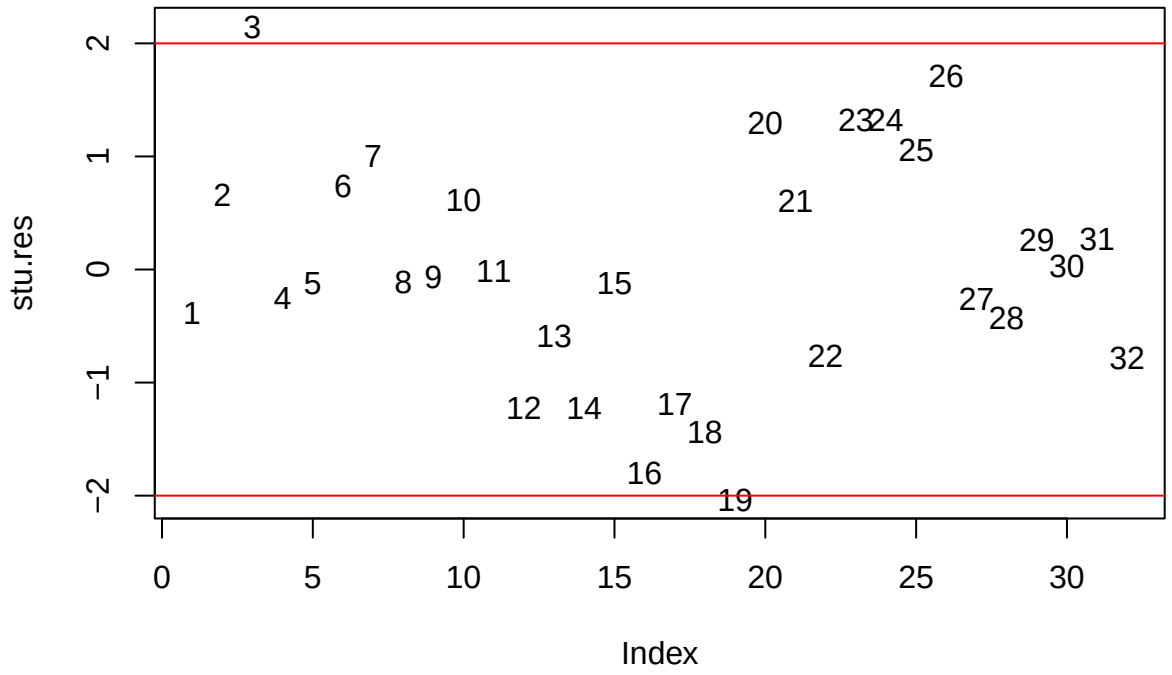


Looking at the plot above, we can confirm that **there are two leverage points, of which county numbers are 31 and 32**.

Detecting outliers

Outliers are observations that are far from the regression surface. We will diagnose outliers by identifying observations with studentized residuals less than -2 or greater than 2.

```
stu.res<-studres(Model1)
plot(stu.res, type="n")
text(stu.res, label=(1:length(county)))
abline(h=2, col="red")
abline(h=-2, col="red")
```



Looking at the plot above, we can conclude that **observations, such as county 3 and county 19, are outliers**.

We can observe that there are no observations that are both leverage points and outliers. In other words, there are no observations sufficient enough to influence the least squares estimates.

Detecting influence

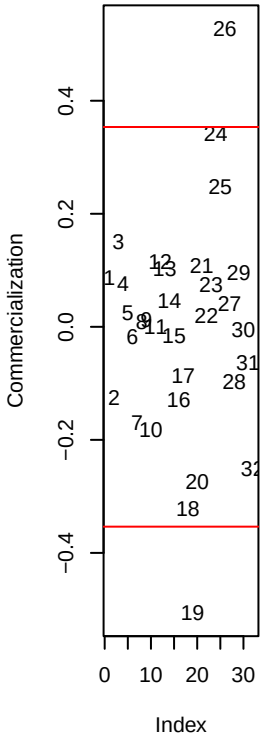
Influential observations occur when the regression results change significantly as a consequence of omitting the observations. We will use DFBETAS to detect them.

```
df<-dfbetas (Model1)
df
```

```
##      (Intercept)    commercial     tradition       peasant       inequal
## 1   0.0336849593  8.739515e-02 -0.0551571057  0.0298494047  0.0162441897
## 2  -0.1223900418 -1.249456e-01  0.1041857928  0.0453659107  0.1266504168
## 3   0.3895998857  1.501679e-01 -0.5152596995  0.2355922031  0.2541393755
## 4   0.1039068615  7.648534e-02 -0.0949135194 -0.0385325047 -0.0736786155
## 5   0.0131288075  2.470434e-02 -0.0084773948 -0.0132971952 -0.0249544984
## 6   0.1323830136 -1.790835e-02 -0.1032163506 -0.1035459218 -0.0803586599
## 7   0.1736218636 -1.694717e-01 -0.0585389977 -0.2654682116 -0.2669556695
## 8  -0.0088869135  8.508493e-03  0.0079789109  0.0052773627 -0.0008745203
## 9   0.0125335340  1.214858e-02 -0.0125036762 -0.0025287794 -0.0076317059
## 10 -0.0490866958 -1.827072e-01  0.0868108159 -0.0387028167 -0.0246683183
## 11 -0.0003735562  9.144823e-05  0.0007025135 -0.0006145865 -0.0010003256
## 12  0.0833807014  1.158921e-01  0.0079320562 -0.1866500976 -0.3231231138
## 13  0.1517269324  1.028541e-01 -0.1400155769 -0.0588789547 -0.1049803865
## 14 -0.0002104415  4.622509e-02 -0.0355972934  0.0777666848  0.0541852480
## 15 -0.0452528506 -1.530805e-02  0.0572788336 -0.0005112614 -0.0227997734
## 16  0.1199552852 -1.282286e-01 -0.0594171031 -0.1044357762 -0.1906547357
## 17 -0.1331359777 -8.638372e-02  0.0539530296  0.2219822630  0.2329015266
## 18 -0.4454318243 -3.226230e-01  0.5498787938  0.0023761251 -0.1390613724
## 19 -0.1993360930 -5.059827e-01  0.1655913604  0.2181339064  0.2235836740
```

```
## 20  0.5773240868 -2.744141e-01 -0.5005380747 -0.3377528185 -0.1987043395
## 21 -0.1263143594  1.082440e-01  0.1274180745  0.0243355411  0.0038203961
## 22 -0.0368677564  2.055798e-02  0.0436978985  0.0287246818 -0.0371400150
## 23  0.1717597183  7.386232e-02 -0.2922612355  0.1147198119  0.3040940292
## 24 -0.3754773743  3.417303e-01  0.3469222548  0.1067076633  0.0833053221
## 25 -0.2523786371  2.480037e-01  0.2486517401  0.0359971023  0.0134657730
## 26  0.0766613905  5.274567e-01 -0.0635783750 -0.0996657310 -0.1625715616
## 27 -0.0326717392  4.205865e-02 -0.0006200121  0.0671679994  0.0770268968
## 28 -0.0243176742 -9.657762e-02  0.0060056584  0.0515925051  0.0717593993
## 29  0.0034493820  9.661200e-02  0.0124073007 -0.0370674954 -0.0651128558
## 30  0.0007214089 -4.240369e-03  0.0061216622 -0.0110893996 -0.0165076543
## 31 -0.0839605692 -6.279982e-02  0.0478604119  0.3198203454  0.1180967302
## 32 -0.2340267325 -2.505985e-01  0.1570624200 -0.2637778478  0.3488979874
```

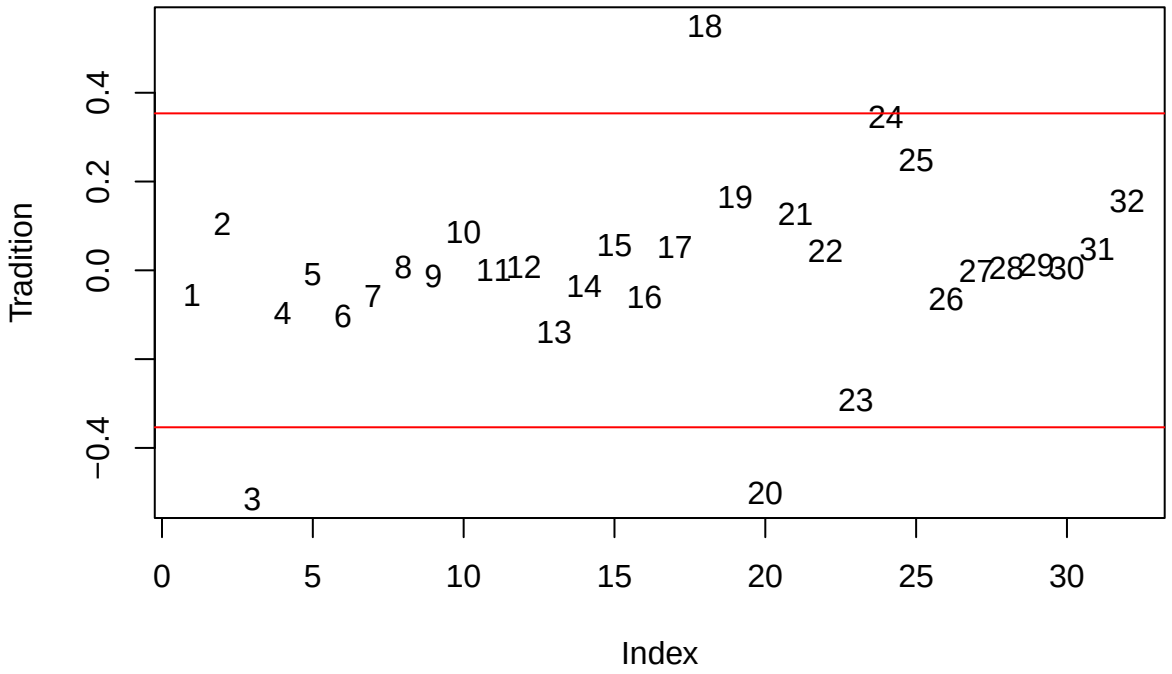
```
df.commercial <-df[,2]
df.tradition <- df[,3]
df.peasant <- df[,4]
df.inequal <- df[,5]
par(mfrow=c(1,4))

#dfbetas for commercialization
plot(df.commercial, type="n", ylab="Commercialization")
text(df.commercial, label=(county))
abline(h=(2/sqrt(n)), col="red")        # approx. cutoff for dfbetas = abs value of 2/sqrt(n)
abline(h=-(2/sqrt(n)), col="red")
```



Looking at the DFBETAS for commercialization, county 19 and 26 are influential observations.

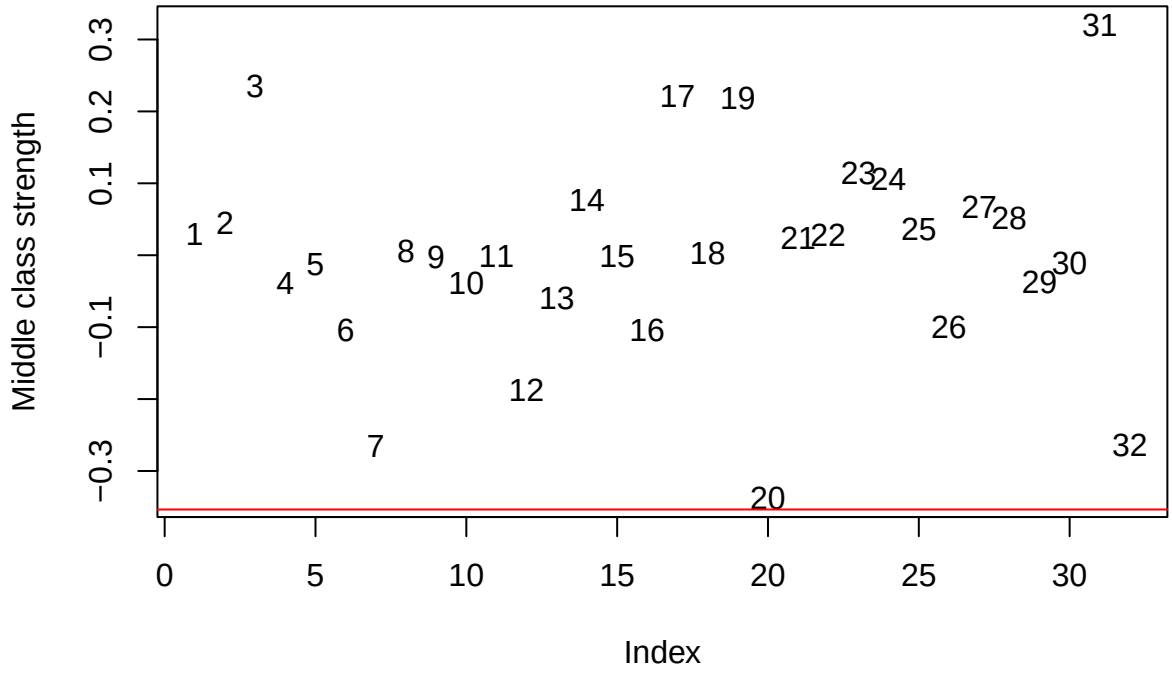
```
#dfbetas for tradition
plot(df[,3], type="n", ylab="Tradition")
text(df[,3], label=(county))
abline(h=(2/sqrt(n)), col="red")
```

```
abline(h=-(2/sqrt(n)), col="red")
```



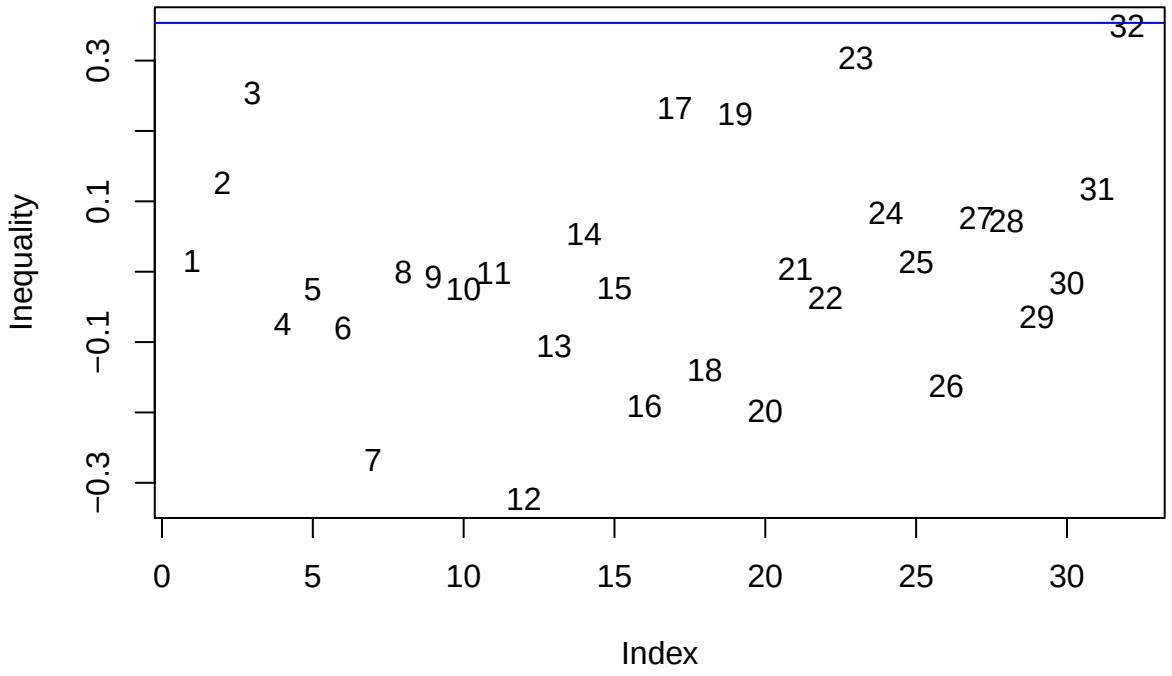
Looking at DFBETAS for traditionalism, county 3, 18, 30 are influential observations.

```
#dfbetas for middle class strength
plot(df[,4], type="n", ylab="Middle class strength")
text(df[,4], label=(county))
abline(h=(2/sqrt(n)), col="blue")
abline(h=-(2/sqrt(n)), col="red")
```



In regards with DFBETAS for middle class strength, there are no influential observations.

```
#dfbetas for inequality
plot(df[,5], type="n", ylab="Inequality")
text(df[,5], label=(county))
abline(h=(2/sqrt(n)), col="blue")
abline(h=-(2/sqrt(n)), col="red")
```



Similarly, there are no influential observations in regards with DFBETAS for inequality.

We will try try fitting model without influential observations (county# 3,18,19,26,30)

```
Model2 <- update(Model1, subset=-c(3,18,19,26,30))
summary(Model2)
```

```
##
## Call:
## lm(formula = intensity ~ commercial + tradition + peasant + inequal,
##     data = chirot, subset = -c(3, 18, 19, 26, 30))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05955 -0.55621 -0.07775  0.78989  1.48869
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.098451   5.088085  -2.378   0.0265 *
## commercial    0.094218   0.019676   4.788 8.81e-05 ***
## tradition     0.110087   0.059599   1.847   0.0782 .
## peasant      -0.008836   0.016090  -0.549   0.5884
## inequal       0.675528   2.777941   0.243   0.8101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 22 degrees of freedom
## Multiple R-squared:  0.6871, Adjusted R-squared:  0.6302
```

```
## F-statistic: 12.08 on 4 and 22 DF,  p-value: 2.408e-05
```

In the new model, the directions of associations remain the same. Furthermore, commercialization is still only an independent variable that is statistically significant. The noticeable change is that the fit of model has improved greatly after omitting the influential observations. Thus, it is reasonable to use the modified model instead of the original one.

Collinearity

We will further examine if there are collinearlity among the independent variables.

```
vif(Model2)
```

```
## commercial  tradition    peasant    inequal
##   1.270428   1.265468   2.189706   2.269151
```

By checking VIF's, it is hard to suspect that any collinearity among the independent variables exist.

Addressing influential data and collinearity

Based on what we found in regression diagnostics, we will address the model by ommitting influential observations. Therefore, I will advise to use Model 2.