Homework 4 - Coding

John L Kaspers

June 3, 2020

Warning

Warning: You must have the “alr4” package, “faraway” package, “car” package, and “carData” package installed for this assignment. You can install these packages by running the following code in base R (not RStudio or RMarkdown).

> install.packages(“alr4”)
> install.packages(“faraway”)
> install.packages(“car”)
> install.packages(“carData”)

Predictor Considerations (9 points)

Fit the requested models for each of the following 3 data sets. Do not include any interaction terms, polynomial terms, or transformations. You will then determine if the models are overparameterized, if collinearity is present (using the `vif()` function), or if the model is overfit. Use the guidelines presented in the lecture notes.

Note: If a model is overparameterized, the `vif()` function will return an error. You must refit a smaller (non-overparameterized) model and then pass that through the function.

1. Using the `stopping` data set in the `alr4` package, create a regression model that predicts `Distance` using `Speed`.

```
data(stopping, package = "alr4")

str(stopping)
```

```
## 'data.frame':    62 obs. of  2 variables:
##  $ Speed   : int  4 5 5 5 5 7 7 8 8 8 ...
##  $ Distance: int  4 2 4 8 8 7 7 8 9 11 ...
```

```
View(stopping)

lm.predict.distance <- lm(Distance~Speed, data = stopping)
summary(lm.predict.distance)
```

```
##
## Call:
## lm(formula = Distance ~ Speed, data = stopping)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.410  -7.343  -1.334   5.927  35.608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.1309     3.2308  -6.231 5.04e-08 ***
## Speed         3.1416     0.1514  20.751  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.77 on 60 degrees of freedom
## Multiple R-squared:  0.8777, Adjusted R-squared:  0.8757
## F-statistic: 430.6 on 1 and 60 DF,  p-value: < 2.2e-16
```

a. Is the model overparameterized? Provide a brief explanation or numerical support with your answer.

No, it is impossible to have an overparametrized model when we just have one predictor.

b. Is there collinearity? Provide a brief explanation or numerical support with your answer.

It’s impossible to have collinearity because we only have one predictor so there would be no correlation between predictors.

c. Is the model overfit? Provide a brief explanation or numerical support with your answer.

No, overfitting can occur due to excess predictors and in this scenario we just have one. By a rule of thumb, n >= 10p, so here n = 60 and p = 1 and we have 60 >= 10(1). Thus the model is not overfit.

---

2. Using the `Bfox` data set in the `carData` package, create a regression model that predicts `menwage` using all other variables.

```
data(Bfox, package = "carData")

#Finding what variables to use
str(Bfox)
```

```
## 'data.frame':    30 obs. of  6 variables:
##  $ partic  : num  25.3 24.4 24.2 24.2 23.7 24.2 24.1 23.8 23.6 24.3 ...
##  $ tfr     : int  3748 3996 3725 3750 3669 3682 3845 3905 4047 4043 ...
##  $ menwage : num  25.4 26.1 25.1 25.4 26.8 ...
##  $ womwage : num  14.1 14.6 14.2 14.6 15.3 ...
##  $ debt    : num  18.2 28.3 30.6 35.8 38.4 ...
##  $ parttime: num  10.28 9.28 9.51 8.87 8.54 ...
```

```
lm.predict.menwage <- lm(menwage ~ partic + tfr + womwage + debt + parttime, data = Bfox)
summary(lm.predict.menwage)
```

```
##
## Call:
## lm(formula = menwage ~ partic + tfr + womwage + debt + parttime,
##     data = Bfox)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09058 -0.53100 -0.01057  0.46169  1.56610
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.4259163  7.8947015   1.067  0.29646
## partic      -0.2094309  0.3809684  -0.550  0.58758
## tfr          0.0018308  0.0006049   3.026  0.00583 **
## womwage      0.8068242  0.2543778   3.172  0.00411 **
## debt         0.0804993  0.0306861   2.623  0.01490 *
## parttime     0.2084427  0.2989035   0.697  0.49228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9812 on 24 degrees of freedom
## Multiple R-squared:  0.9842, Adjusted R-squared:  0.9809
## F-statistic: 298.8 on 5 and 24 DF,  p-value: < 2.2e-16
```

```
# Is model collinear?
vif(lm.predict.menwage)
```

```
##     partic        tfr    womwage       debt   parttime
## 154.777814   5.863347  43.049163 108.962068  30.137304
```

a. Is the model overparameterized? Provide a brief explanation or numerical support with your answer.

No, the model is not overparameterized because when running the regression I did not get any NA values in the output. This is assuming menwage and womwage are not factor variables.

b. Is there collinearity? Provide a brief explanation or numerical support with your answer.

Yes, there is collinearity with partic, womwage, debt, and parttime which all have vif values well above 10.

c. Is the model overfit? Provide a brief explanation or numerical support with your answer.

Overfitting occurs due to excess predictors and by a rule of thumb, n >= 10p, so here n = 30. 30 >= 10(5) -> 30 is not greator than or equal to 50 so the model is overfit.

---

3. Using the `sat` data set in the `faraway` package, create a regression model that predicts `salary` using all other variables.

```
data(sat, package = "faraway")

#Determining "all other variables"
str(sat)
```

```
## 'data.frame':    50 obs. of  7 variables:
##  $ expend: num  4.41 8.96 4.78 4.46 4.99 ...
##  $ ratio : num  17.2 17.6 19.3 17.1 24 18.4 14.4 16.6 19.1 16.3 ...
##  $ salary: num  31.1 48 32.2 28.9 41.1 ...
##  $ takers: int  8 47 27 6 45 29 81 68 48 65 ...
##  $ verbal: int  491 445 448 482 417 462 431 429 420 406 ...
##  $ math  : int  538 489 496 523 485 518 477 468 469 448 ...
##  $ total : int  1029 934 944 1005 902 980 908 897 889 854 ...
```

```
lm.predict.salary <- lm(salary ~ expend + ratio + takers + verbal + math, data = sat)
summary(lm.predict.salary)
```

```
##
## Call:
## lm(formula = salary ~ expend + ratio + takers + verbal + math,
##     data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2839 -1.0490 -0.0523  0.9140  4.2697
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.994881  10.638677  -1.127    0.266
## expend        3.908133   0.308091  12.685 2.69e-16 ***
## ratio         0.975425   0.145998   6.681 3.36e-08 ***
## takers        0.050735   0.030015   1.690    0.098 .
## verbal       -0.005259   0.039703  -0.132    0.895
## math          0.015562   0.032178   0.484    0.631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.052 on 44 degrees of freedom
## Multiple R-squared:  0.8929, Adjusted R-squared:  0.8807
## F-statistic: 73.33 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
#Is model collinear?
vif(lm.predict.salary)
```

```
##    expend     ratio    takers    verbal      math
##  2.050826  1.273655  7.506217 22.690255 19.470915
```

a. Is the model overparameterized? Provide a brief explanation or numerical support with your answer.

Yes, the model is overparameterized because R gives NA values for total in the summary, thus total = math + verbal. Because of this, I must remove total to have a better model.

b. Is there collinearity? Provide a brief explanation or numerical support with your answer.

Yes there is collinearity with verbal and math variables, as shown by both having VIF values above 10. The vif value for takers (7.5) is also high.

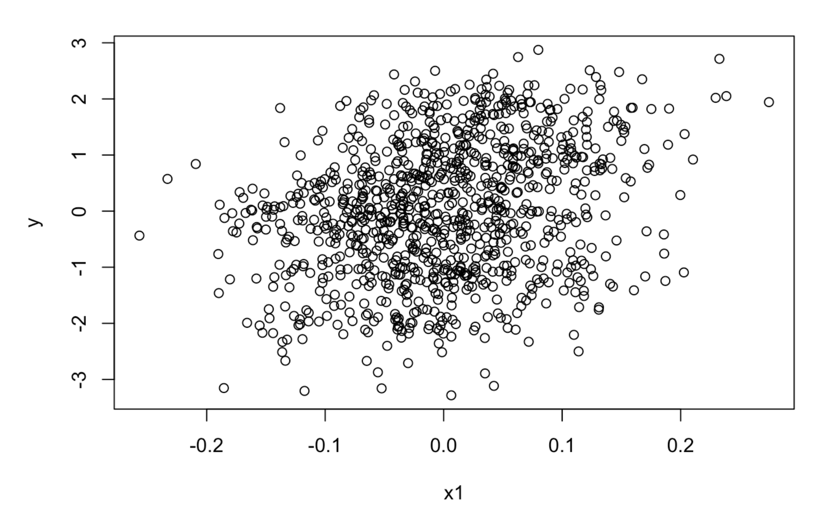
c. Is the model overfit? Provide a brief explanation or numerical support with your answer.

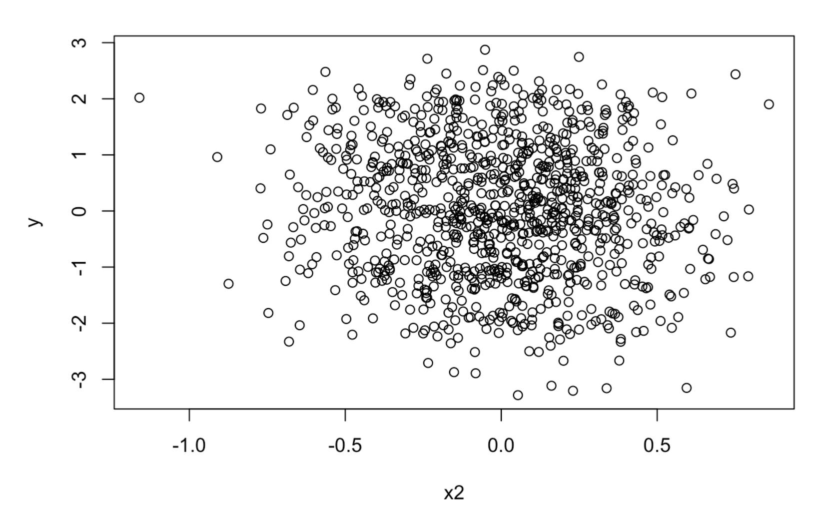
Well that depends, using all of the (6) predictors violates our rule of thumb of n >= 10p because there are 50 observations and 50 is not >= 10*6, so using all predictors creates an overfit model. If however I remove ‘total’ as a predictor, then I have 50 >= 50 and the model is no longer overfit.

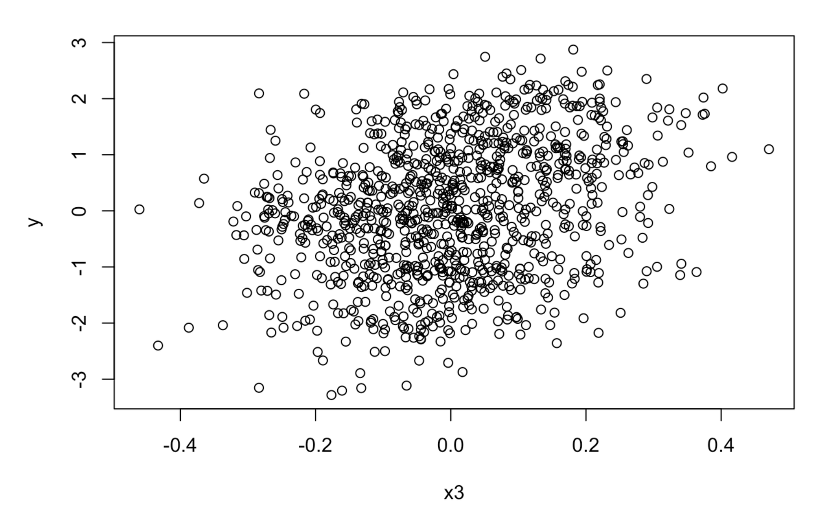
Regression Diagnostics - Part I (3 points)

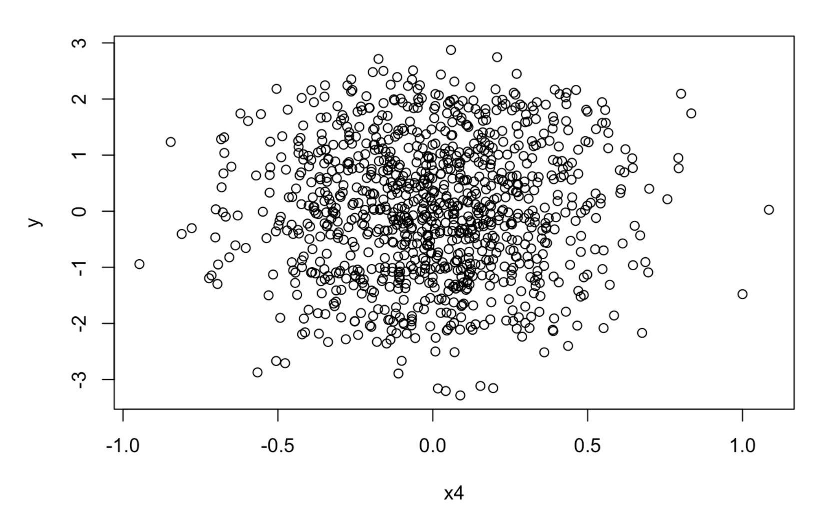
1. Using the `Rpdata` data set in the `alr4` package, create a scatterplot matrix of all data (`y`, `x1`,..., `x6`). Note: the variables are artifical and have no interpretable meaning.

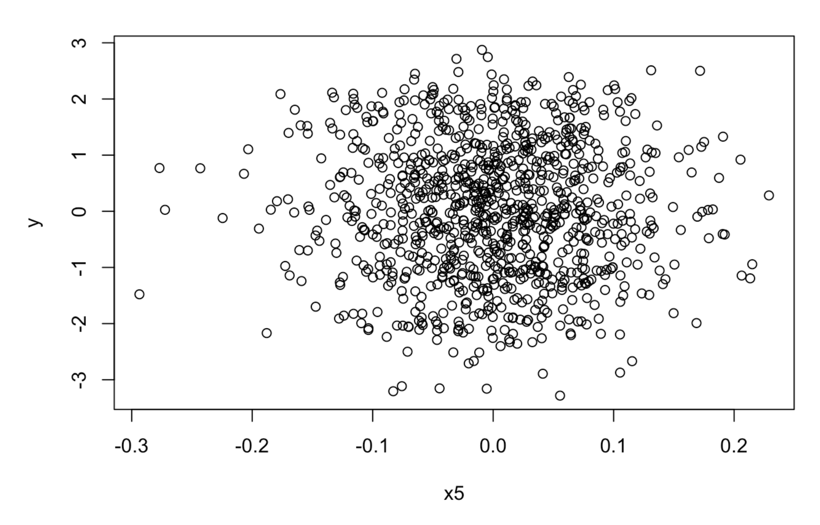
```
data(Rpdata, package = "alr4")

plot(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
```

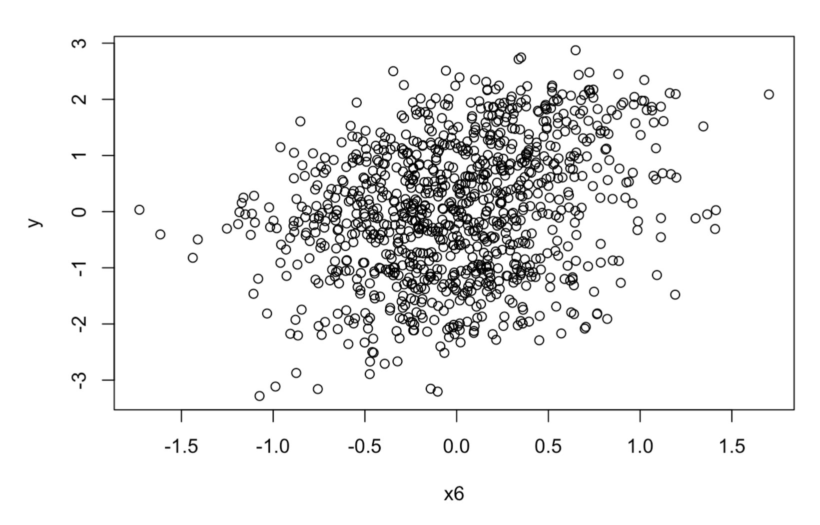












Do any of the individual predictors have a non-linear relationship with the response?

No, all of the individual predictors appear to follow MVN distributions. They’re all approximately linear.

2. Create a regression model that predicts `y` using all other variables (`x1`,..., `x6`). Pass your model through the `summary()` function.

```
lm.predict.y <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
summary(lm.predict.y)
```

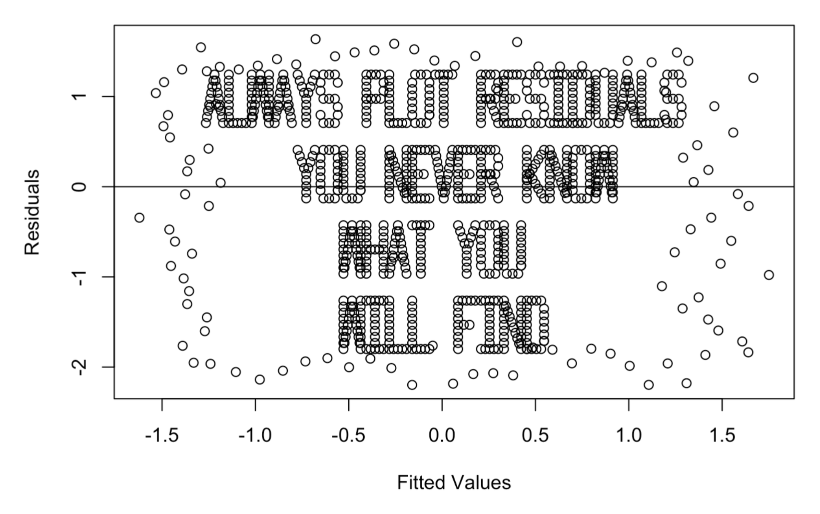
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1977 -0.7631  0.1729  0.8851  1.6359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02481    0.03188   0.778    0.437
## x1           4.14061    0.50954   8.126 1.32e-15 ***
## x2           1.01233    0.15522   6.522 1.11e-10 ***
## x3           3.99614    0.32663  12.234  < 2e-16 ***
## x4           0.96045    0.16657   5.766 1.09e-08 ***
## x5           3.75122    0.64726   5.796 9.17e-09 ***
## x6           0.95390    0.08561  11.142  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.003 on 983 degrees of freedom
## Multiple R-squared:  0.3112, Adjusted R-squared:  0.307
## F-statistic: 74.03 on 6 and 983 DF,  p-value: < 2.2e-16
```

Are any of the predictors statistcally significant (at a 5% level)?

Yes, x1, x2, x3, x4, x5 and x6 are all statistically significant at the 5% level.

3. Create a scatterplot of the residuals versus the fitted values.

```
# Write your code here
lm.model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
plot(lm.model$fitted.values, lm.model$residuals,
      xlab = "Fitted Values",
      ylab = "Residuals")
abline(h=0)
```



What is the message of this analysis?

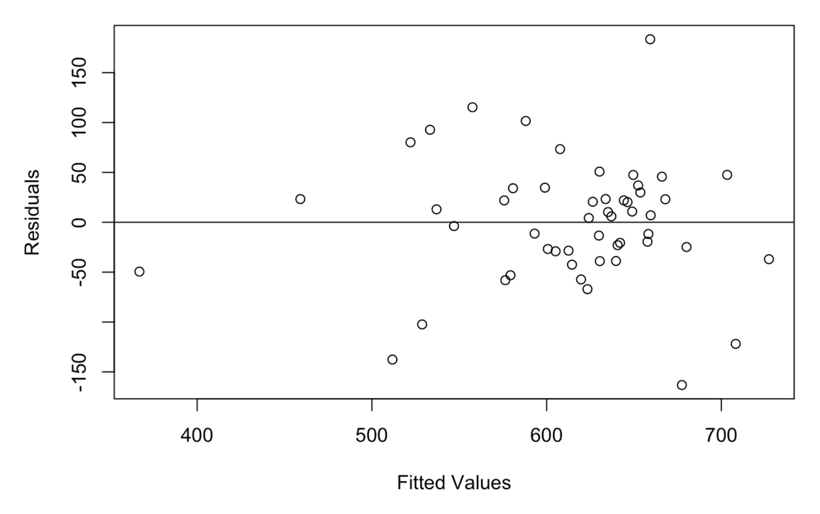
Always plot residuals you never know what you will find.

Regression Diagnostics - Part II (4 points)

Recall the fuel data from several lectures back. We will take the model we analyzed and check the regression diagnostics. The model has been created for you as `lm.fuel`.

1. Using the `lm.fuel` model above, create a plot of the residuals versus fitted values. Be sure to include appropriate labels for your plot axes.

```
lm.model <- lm(Fuel ~ ., data = fuel2001)
plot(lm.model$fitted.values, lm.model$residuals,
      xlab = "Fitted Values",
      ylab = "Residuals")
abline(h=0)
```



2. Are the zero-mean assumption and constant variance assumption reasonably met? Briefly explain why or why not.
#It appears that the constant variance assumption is violated - the variance is small and # gets larger as we move along the fitted values

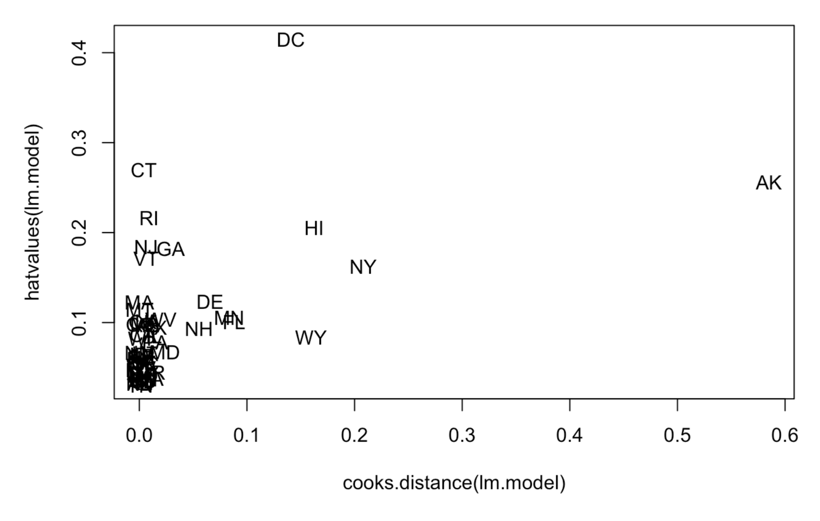
Yes, the zero-mean assumption and constant variance assumpmtion are reasonably met, as shown by the plot which has approximately a mean residual of zero and variance approximately around zero.

3. Create a plot of leverage versus Cook’s distance. Be sure to include appropriate labels for your plot axes. You may utilize the `hatvalues()` function to find the leverages and the `cooks.distance()` function to find the values for Cook’s distance.

Note: The following code structure will allow you to add the state names to the plot (instead of just plotting the points).

> plot(x, y, col = “white”)
> text(x, y, labels = rownames(dataset))

```
plot(cooks.distance(lm.model), hatvalues(lm.model), col = "white")
text(cooks.distance(lm.model), hatvalues(lm.model), labels = rownames(fuel2001))
```



4. Which location has the highest value for Cook’s distance? Which location has the highest leverage? According to our rule of thumb for Cook’s distance, do we have any ‘problematic outliers’?

AK has the highest value for Cook’s distance. DC has the highest leverage. According to our rule of thumb for Cook’s distance, AK has a high (>0.5) Cook’s distance, but no we do not have any ‘problematic outliers’ (Cook’s distance > 1).