Anagha Uppal - MW Section - Homework 6

Multiple Regression 1

due Oct 6 by 330

Question 1

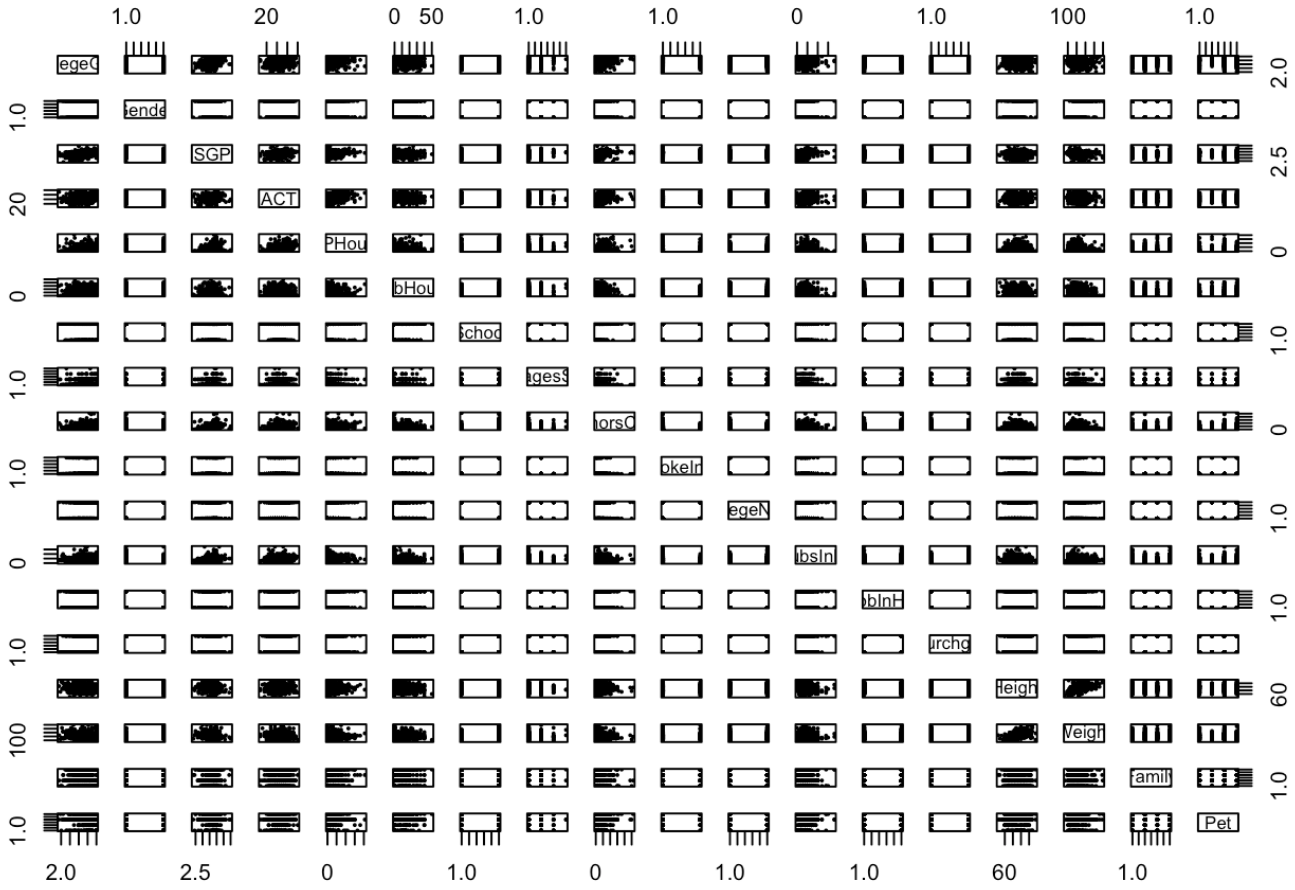
Load in the `EDUCATION` dataset using the `data` command. This is the education analytics example where we are interested in studying what quantities that are known about students in high school have an association with college "success" as measured by `CollegeGPA`.

```
#Code reading in EDUCATION dataset
data("EDUCATION")
```

The quantitative predictors of interest in the dataset are `HSGPA`, `ACT`, `APHours`, `HSHonorsClasses`, `ClubsInHS`.

a. Provide a scatterplot matrix for the `CollegeGPA` and the 5 predictors mentioned above. Add the additional arguments `pch=20` and `cex=0.3` to make the points small. Are there any nonlinearities in the relationships between `CollegeGPA` and the 5 predictors that may prevent the multiple regression model from fitting the relationships?

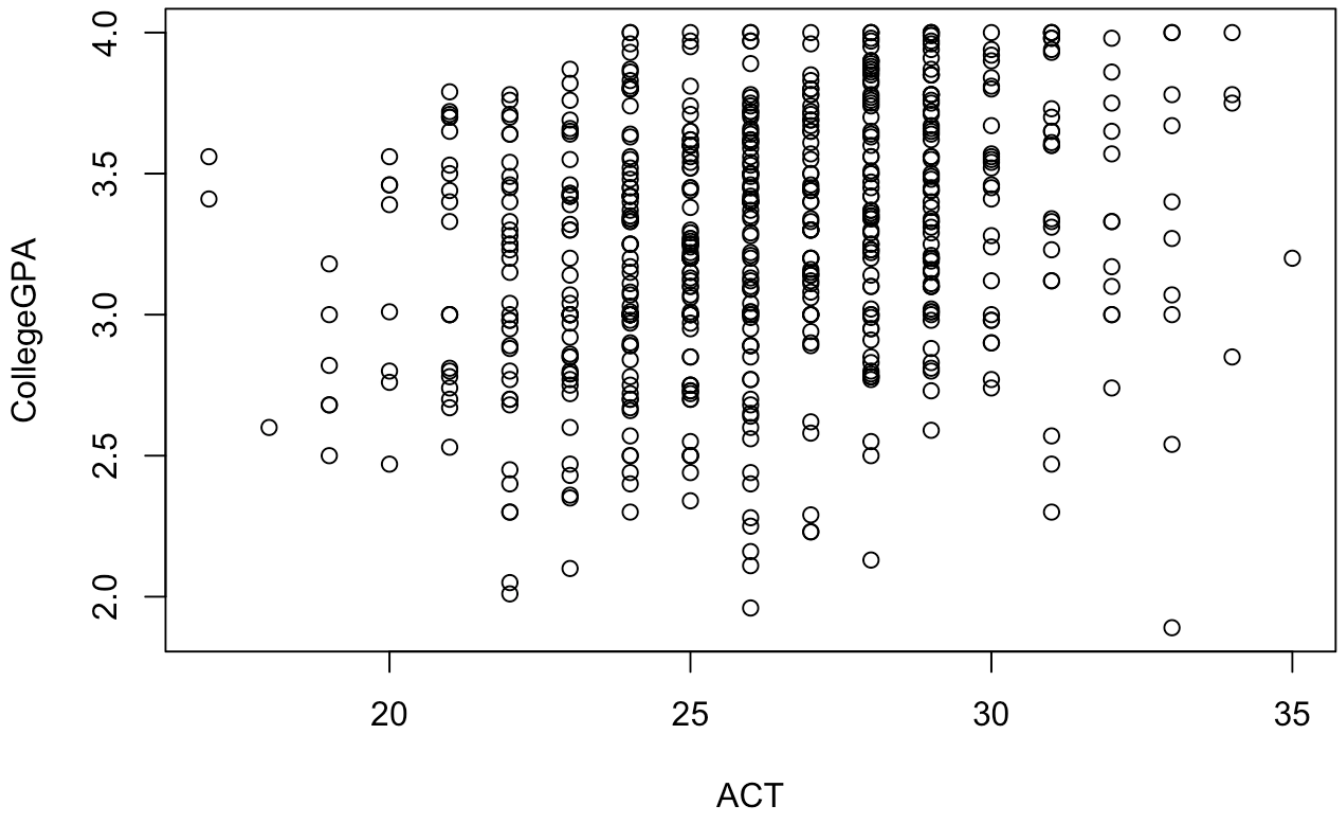
```
#Scatterplot matrix
pairs(CollegeGPA~.,data=EDUCATION,pch=20,cex=0.3)
```



Response: I cannot visibly identify any striking nonlinearities in the matrix of scatterplots.

b. Fit a simple linear regression predicting `CollegeGPA` from `ACT`. If two students differed in ACT score by 5 points, by how much do we expect their College GPAs to differ (and who is expected to have the higher GPA)?

```
#Code fitting model predicting CollegeGPA from ACT and printing out the summary
plot(CollegeGPA~ACT, data=EDUCATION)
```




```
M <- lm(CollegeGPA~ACT,data=EDUCATION)
summary(M)
##
## Call:
## lm(formula = CollegeGPA ~ ACT, data = EDUCATION)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62891 -0.29598  0.02378  0.35378  0.80597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.327668   0.149379  15.582  < 2e-16 ***
## ACT         0.036098   0.005675   6.361 3.96e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4337 on 605 degrees of freedom
## Multiple R-squared:  0.06268,    Adjusted R-squared:  0.06113
## F-statistic: 40.46 on 1 and 605 DF,  p-value: 3.96e-10
```

Response: Model is CollegeGPA=0.036098*ACT + 2.327668 The model predicts a student whose ACT score is 5 points higher than another's will also score a College GPA that is 0.18049 higher than the other's.

c. Fit the remaining 4 simple linear regressions (predicting `CollegeGPA` from each variable separately). Report which variables had statistically significant regressions.

```
#Note:  leave eval=FALSE above so that the code is included but not the output
#Code for the simple linear regressions
hsgpa <- lm(CollegeGPA~HSGPA,data=EDUCATION)
summary(hsgpa) # = 2.2e-16; statistically significant
ap <- lm(CollegeGPA~APHours,data=EDUCATION)
summary(ap) # = 0.00127; statistically significant

honors <- lm(CollegeGPA~HSHonorsClasses,data=EDUCATION)
summary(honors) # = 2.221e-06; statistically significant

clubs <- lm(CollegeGPA~ClubsInHS,data=EDUCATION)
summary(clubs) # = 0.0108; statistically significant
```

Significant: ALL - HSGPA, APHours, HonorsClasses, ClubsInHS

Non-significant: NONE

d. Fit the multiple regression model predicting `CollegeGPA` from all 5 variables simultaneously. Include the summary.

```
#Code fitting multiple regression model
allreg <- lm(CollegeGPA~ACT+HSGPA+APHours+HSHonorsClasses+ClubsInHS, data=EDUCATIO
N)
#Code for summary
summary(allreg)
##
## Call:
## lm(formula = CollegeGPA ~ ACT + HSGPA + APHours + HSHonorsClasses +
##     ClubsInHS, data = EDUCATION)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63760 -0.26308  0.03337  0.31280  1.07687
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.1628768  0.2145623   5.420 8.65e-08 ***
## ACT              0.0225991  0.0058399   3.870 0.000121 ***
## HSGPA            0.4068052  0.0514215   7.911 1.23e-14 ***
## APHours         -0.0006344  0.0027229  -0.233 0.815845
## HSHonorsClasses  0.0029927  0.0026097   1.147 0.251933
## ClubsInHS       -0.0025397  0.0070636  -0.360 0.719308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4099 on 601 degrees of freedom
## Multiple R-squared:  0.1684, Adjusted R-squared:  0.1614
## F-statistic: 24.34 on 5 and 601 DF,  p-value: < 2.2e-16
```

- The typical size of the residual for this model is: 0.03337
- The value of R^2 is: 0.1684
- The value of R^2_{adj} is: 0.1614

e. Interpretation of the coefficients in a multiple regression model is a key skill I want you to acquire in this class.

- Provide a precise and detailed interpretation of the coefficient of ACT

Response: If two individuals have otherwise identical data, but they differ only in their ACT score by a point, the person with the higher ACT score is expected to earn a CollegeGPA that is higher by 0.0225991.

- Provide a precise and detailed interpretation of the coefficient of ClubsInHS

Response: If two individuals have otherwise identical data, but they differ only in the number of clubs they participated in in high scool by one club, the person with the higher club number is expected to earn a CollegeGPA that is lower by 0.0025397.

f. Interpreting statistical significance is another key skill I want you to acquire in this class.

- ACT is statistically significant in the multiple regression model. What precisely does this mean?

Response: This means that the ACT predictor still adds to the provision of valuable information to predict CollegeGPA. When ACT is added to the model containing all the predictors, the reduction in SSE is large, which means that the result was unlikely to have been produced by chance.

- ClubsInHS is NOT statistically significant in the multiple regression model but it is in the simple linear regression. As if you were talking to your boss, explain this apparent contradiction and address whether ClubsInHS does or does not contain any information about CollegeGPA.

Response: ClubsInHS by itself offers valuable information and is able to predict CollegeGPA fairly well. However, when combined with various other predictor variables, it does not add very significant information that impacts the model. Therefore, it is not considered statistically significant as part of the multiple regression model.

- True or False (and explain): since APHours is not a statistically significant predictor in the multiple regression model, you should recommend to high school freshmen to skip any AP classes since they won't help you do any better in college.

Response: False. APHours is an excellent predictor of college success. However, involvement in AP classes tends to coincide with the impact on the model by variables like HSGPA and HSHonorsClasses. Therefore, it is not considered statistically significant as it does not add unique information that impacts the model. I would not recommend that the freshmen skip AP classes.

g. If college admissions counselors were interested in studying the association between college GPA and ACT score, why should they be looking at its coefficient in the multiple regression model and not at the coefficient in the simple linear regression?

Response: When we study a model using our predictors one at a time, the other predictor variables, still causing the model to be affected, act as lurking variables and cause our conclusions to sometimes be invalid. We must study a multiple regression model to have the full, big picture at once for accuracy.

h. Why in general do we refer to R^2_{adj} instead of R^2 when talking about how well a model fits the data?

Response: As we increase the number of predictor variables for a multiple regression model, the R^2 will continue to increase. Therefore, there must be a "penalty" to limit the number of predictor variables used to the minimum most meaningful. The higher the number of predictor variables, the greater a portion is subtracted from the R^2 to result in the adjusted R^2 value.

Question 2

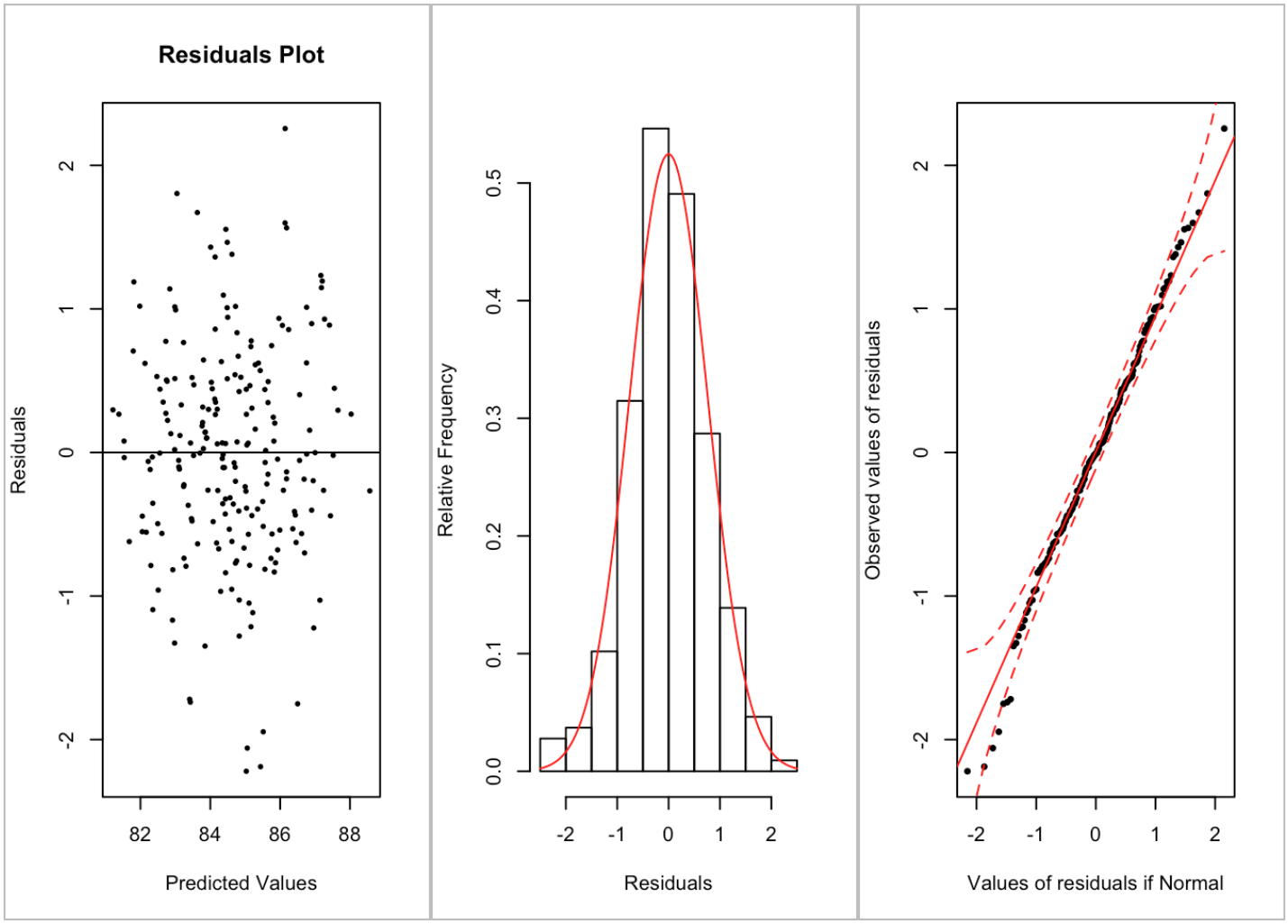
Load in `EX4.STOCKS` using the `data` command. The goal is to predict the closing price of Alcoa stock (`AA`) from the closing prices of other stocks and commodities two days prior (`IMBlag2`, `HongKonglag2`, etc.). If any associations between the prices continued into the future, you would be able to leverage this information to make a fortune.

```
#Code loading in `EX4.STOCKS`
data("EX4.STOCKS")
```

a. Fit the multiple regression model predicting `AA` from every other variable using the "twidle dot" shortcut. Run `check.regression` (do not add the argument `extra=TRUE` here) and examine the

results.

```
#Code fitting the model predicting AA from all variables
stocksreg <- lm(AA~., data=EX4.STOCKS)
#Code for check.regression
check.regression(stocksreg)
```



```
##
## Tests of Assumptions: ( sample size n = 216 ):
## Linearity
##    p-value for AAPLlag2 : 0.6573
##    p-value for AXPlag2 : 0.9785
##    p-value for BAlag2 : 0.6807
##    p-value for BAClag2 : 0.7448
##    p-value for CATlag2 : 0.8374
##    p-value for CSCOlag2 : 0.3185
##    p-value for CVXlag2 : 0.09
##    p-value for DDlag2 : 0.1123
##    p-value for DISlag2 : 0.0012
##    p-value for GElag2 : 0.4561
##    p-value for HDlag2 : 0.12
##    p-value for HPQlag2 : 0.5096
##    p-value for IBMlag2 : 0.4367
##    p-value for INTClag2 : 0.0881
##    p-value for JNJlag2 : 0.0136
##    p-value for JPMlag2 : 0.7806
##    p-value for KOlag2 : 0.4892
##    p-value for MCDlag2 : 0.9947
##    p-value for MMMlag2 : 0.3225
##    p-value for MRKlag2 : 0.0955
##    p-value for MSFTlag2 : 0.7247
##    p-value for PFElag2 : 0.6895
##    p-value for PGlag2 : 0.904
##    p-value for Tlag2 : 0.0128
##    p-value for TRVlag2 : 0.2055
##    p-value for UNHlag2 : 0.0363
##    p-value for VZlag2 : 0.0369
##    p-value for WMTlag2 : 0.0383
##    p-value for XOMlag2 : 0.1445
##    p-value for Australialag2 : 0.0045
##    p-value for Copperlag2 : 0.4528
##    p-value for DollarIndexlag2 : 0.8805
##    p-value for Europelag2 : 0.7677
##    p-value for Exchangelag2 : 0.0999
##    p-value for GlobalDowlag2 : 0.0959
##    p-value for HongKonglag2 : 0.0433
##    p-value for Indialag2 : 0.3954
##    p-value for Japanlag2 : 0.2058
##    p-value for Oillag2 : 0.0058
##    p-value for Shanghailag2 : 0.0439
##    p-value for overall model : NA (not enough duplicate rows)
## Equal Spread:  p-value is 0.2991
## Normality:  p-value is 0.4097
##
## Advice:  if n<25 then all tests must be passed.
## If n >= 25 and test is failed, refer to diagnostic plot to see if violation is
severe
```

```
##  or is small enough to be ignored.
summary(stocksreg)
##
## Call:
## lm(formula = AA ~ ., data = EX4.STOCKS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22053 -0.47839 -0.00788  0.48991  2.25632
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     45.7105487 25.7203654   1.777 0.077270 .
## AAPLlag2        -0.0031481  0.0058065  -0.542 0.588390
## AXPlag2         -0.2695886  0.1066324  -2.528 0.012349 *
## BAlag2           0.2448629  0.0987799   2.479 0.014127 *
## BAClag2         -0.1106793  0.4240450  -0.261 0.794393
## CATlag2          0.0710758  0.0467816   1.519 0.130489
## CSCOlag2         0.3477485  0.2193249   1.586 0.114649
## CVXlag2          0.0209692  0.0738098   0.284 0.776672
## DDlag2           0.0132809  0.1406637   0.094 0.924887
## DISlag2          0.1846809  0.1731609   1.067 0.287654
## GElag2          -1.2917547  0.3630878  -3.558 0.000482 ***
## HDlag2           0.0385060  0.0963074   0.400 0.689774
## HPQlag2         -0.4908839  0.1174179  -4.181 4.59e-05 ***
## IBMlag2         -0.0142440  0.0326644  -0.436 0.663323
## INTClag2        -0.0494776  0.2155422  -0.230 0.818710
## JNJlag2         -0.2791187  0.1570806  -1.777 0.077320 .
## JPMlag2          0.3309126  0.1054818   3.137 0.002002 **
## KOlag2           0.4149904  0.2082417   1.993 0.047834 *
## MCDlag2          0.0006411  0.0663642   0.010 0.992303
## MMMlag2         -0.0738194  0.1165547  -0.633 0.527335
## MRKlag2          0.2244010  0.1910005   1.175 0.241642
## MSFTlag2         0.5247395  0.2282121   2.299 0.022666 *
## PFElag2          0.0540063  0.3462034   0.156 0.876216
## PGlag2          -0.0743877  0.0977481  -0.761 0.447673
## Tlag2            0.1068670  0.2574679   0.415 0.678600
## TRVlag2          0.3205066  0.0999364   3.207 0.001594 **
## UNHlag2         -0.0519570  0.0705112  -0.737 0.462193
## VZlag2           0.0646278  0.1890175   0.342 0.732826
## WMTlag2          0.0593871  0.0745970   0.796 0.427049
## XOMlag2         -0.3068335  0.1300128  -2.360 0.019375 *
## Australialag2    0.0062963  0.0027734   2.270 0.024410 *
## Copperlag2       3.4161253  1.4625189   2.336 0.020636 *
## DollarIndexlag2  0.3691260  0.2160579   1.708 0.089324 .
## Europelag2       0.0311160  0.0425729   0.731 0.465825
## Exchangelag2     0.1390952  0.0746496   1.863 0.064094 .
## GlobalDowlag2   -0.0821246  0.0836923  -0.981 0.327815
## HongKonglag2    -0.0704035  0.0195619  -3.599 0.000416 ***
## Indialag2        0.0042402  0.0021776   1.947 0.053110 .
## Japanlag2       -0.0018522  0.0007007  -2.643 0.008957 **
```

```
## Oillag2          0.1821036  0.0458336   3.973 0.000104 ***
## Shanghailag2    -0.0040485  0.0024759  -1.635 0.103804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8427 on 175 degrees of freedom
## Multiple R-squared:  0.8007, Adjusted R-squared:  0.7552
## F-statistic: 17.58 on 40 and 175 DF,  p-value: < 2.2e-16
```

- The overall test for linearity of the model as a whole is unavailable since no two rows contained exact duplicates of all the predictor variables. Based on the residuals plot, does linearity seem like a reasonable assumption? Explain.

Response: The points on this plot fall approximately equally above and below the horizontal axis. They are spread randomly and there appear no visible patterns in the graph.

- With so many predictors, we expect a few of the tests for linearity to be failed just by bad luck. If we consider the linearity test to be failed only if the p-value is less than 0.005, which two variables fail the test?

Response: DISlag2 and Australialag2 would fail the test for linearity.

- Rerun `check.regression` but with the command `extra=TRUE` to look at the predictor vs. residual plots of the two failures. Do you believe they are deal-breaking violations? Explain.

```
#Note:  leave eval=FALSE above so the code is shown but not the billion plots
#Code rerunning check.regression
check.regression(stocksreg, extra=TRUE)
```

Response: These graphs are almost perfect - there appear to be no violations here.

- Are the equal spread and Normality assumptions reasonable here? Explain.

Response: Yes, they are both reasonable.

b. As a whole, stock and commodity prices are often strongly correlated with one another. Highly correlated predictors can cause complications. Run `VIF` on the model (make sure you have run lines 12-42 here so that the function is defined).

```
#Code running VIF
VIF(stocksreg)
##        AAPLlag2         AXPlag2          BAlag2         BAClag2         CATlag2
##       54.283373       33.076468       11.965351       45.694081       83.866158
##        CSCOlag2         CVXlag2          DDlag2         DISlag2          GElag2
##       32.844017       42.970842       21.598019      169.718544       60.961576
##          HDlag2         HPQlag2         IBMlag2        INTClag2         JNJlag2
##       73.448560       73.177784       17.051200       46.516556       36.063045
##         JPMlag2          KOlag2          MCDlag2         MMMlag2         MRKlag2
##       44.468613       45.263752       28.039331       40.234695      102.312489
##        MSFTlag2         PFElag2          PGlag2           Tlag2         TRVlag2
##       25.656592       59.106622       17.674597      162.987817       44.939285
##         UNHlag2          VZlag2          WMTlag2         XOMlag2   Australialag2
##        7.770408       90.924463       68.401123       51.474731       27.214321
##      Copperlag2 DollarIndexlag2      Europelag2     Exchangelag2   GlobalDowlag2
##       23.522426       31.428614       57.330160       24.375718      184.226375
##    HongKonglag2       Indialag2       Japanlag2          Oillag2     Shanghailag2
##       34.200493       34.139503       33.586960       36.865753       29.567446
```

- Is multicollinearity a “problem” in this model? Why or why not?

Response: There does not seem to be any problems with multicollinearity in this model. The points are fairly evenly distributed and there is no obvious curvature in the graph.

- In general, when multicollinearity is a “problem”, what are the consquences? In other words, are the values of the coefficients poorly estimated and/or are the predicted values of the model poorly estimated?

Response: Multicollinearity is a “problem” in and of itself. However, it suggests that the standard errors will be really large - the coefficients of the variables provide redundant information.

c. The quality of prediction intervals is extremely sensitive to the assumptions behind the regression being true, but they actually seem to hold pretty well here. Let us make 95% prediction intervals for the closing price of `AA` . As you saw in Activity 6, creating a data frame containing the values of the predictor variables is tedious, so the data is available in `regclass` by loading up `EX4.STOCKPREDICT` using the `data` command.

```
#Code loading in EX4.STOCKPREDICT
data("EX4.STOCKPREDICT")
#Code making 95% prediction intervals for the data in EX4.STOCKPREDICT
predict(stocksreg,newdata=EX4.STOCKPREDICT,interval="prediction")
##        fit      lwr      upr
## 1 85.76773 83.94066 87.59479
## 2 83.78114 81.98734 85.57494
## 3 84.31747 82.54399 86.09095
## 4 85.92866 84.16051 87.69680
## 5 82.16582 80.36011 83.97153
# The prediction interview falls between 84.16051 and 87.69680
```

d. Over the 216 days in which the data was recorded, the closing price of `AA` ranged between 81 and

88 dollars (you can check that with `summary(EX4.STOCKS$AA)`). Further, closing prices on sequential days never differed by more than 1 dollar. With that in mind, do you believe the 95% prediction intervals narrow down the future closing prices enough to be useful and exploited? Explain.

Response: The prediction interval is even broader than the common knowledge of the interval never being over 1 dollar. It provides no new information, and is therefore not useful at all.