Lab 3

Anya Conti

4/22/2017

```
data <- read.csv("http://people.math.umass.edu/~jstauden/outlierdata.csv")
fit <- lm(y~.,data=data,x=T)
```

1.

```
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ ., data = data, x = T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8901 -11.5019  -0.6577  11.8651  27.9840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.6351     1.2578  24.356   <2e-16 ***
## x1            0.9003     3.0797   0.292    0.770
## x2           -4.0704     3.1886  -1.277    0.204
## x3            2.9605     2.9883   0.991    0.323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.28 on 146 degrees of freedom
## Multiple R-squared:  0.01252,    Adjusted R-squared:  -0.007767
## F-statistic: 0.6172 on 3 and 146 DF,  p-value: 0.6049
```

```
fit$x
```

```
##    (Intercept)          x1          x2          x3
## 1            1 0.265508663 0.061464497 0.067371223
## 2            1 0.372123900 0.055715954 0.009485786
## 3            1 0.572853363 0.032877732 0.049259612
## 4            1 0.908207790 0.045313145 0.046155184
## 5            1 0.201681931 0.050044097 0.037521653
## 6            1 0.898389685 0.018086636 0.099109922
## 7            1 0.944675269 0.052963060 0.017635071
## 8            1 0.660797792 0.007527575 0.081343521
## 9            1 0.629114044 0.027775593 0.006844664
## 10           1 0.061786270 0.021269952 0.040044975
## 11           1 0.205974575 0.028479048 0.014114433
## 12           1 0.176556753 0.089509410 0.019330986
## 13           1 0.687022847 0.044623532 0.084135172
## 14           1 0.384103718 0.077998489 0.071991399
## 15           1 0.769841420 0.088061903 0.026721208
## 16           1 0.497699242 0.041312421 0.049500164
## 17           1 0.717618508 0.006380848 0.008311390
## 18           1 0.991906095 0.033548749 0.035388424
```

```
## 19              1 0.380035179 0.072372595 0.096920881
## 20              1 0.777445221 0.033761533 0.062471419
## 21              1 0.934705231 0.063041412 0.066461825
## 22              1 0.212142521 0.084061455 0.031248966
## 23              1 0.651673766 0.085613166 0.040568961
## 24              1 0.125555096 0.039135928 0.099607737
## 25              1 0.267220669 0.038049389 0.085508236
## 26              1 0.386114093 0.089544543 0.095354840
## 27              1 0.013390333 0.064431576 0.081230509
## 28              1 0.382387957 0.074107865 0.078218212
## 29              1 0.869690846 0.060530345 0.026787813
## 30              1 0.340348997 0.090308161 0.076215153
## 31              1 0.482080115 0.029373016 0.098631159
## 32              1 0.599565825 0.019126011 0.029360555
## 33              1 0.493541307 0.088645094 0.039935111
## 34              1 0.186217601 0.050333949 0.081213152
## 35              1 0.827373319 0.087705754 0.007715167
## 36              1 0.668466738 0.018919362 0.036369681
## 37              1 0.794239861 0.075810305 0.044259247
## 38              1 0.107943626 0.072449889 0.015671413
## 39              1 0.723710946 0.094372482 0.058220527
## 40              1 0.411274430 0.054764659 0.097016218
## 41              1 0.820946294 0.071174387 0.098949983
## 42              1 0.647060194 0.038890510 0.017645204
## 43              1 0.782932762 0.010087313 0.054213042
## 44              1 0.553036312 0.092730209 0.038430389
## 45              1 0.529719580 0.028323250 0.067616405
## 46              1 0.789356232 0.059057316 0.026929378
## 47              1 8.000000000 8.000000000 8.000000000
## 48              1 0.477230065 0.084050703 0.017180008
## 49              1 0.732313739 0.031796368 0.036918946
## 50              1 0.692731556 0.078285134 0.072540527
## 51              1 0.047761962 0.267508207 0.048614910
## 52              1 0.086120948 0.218645285 0.006380247
## 53              1 0.043809711 0.516796836 0.078454623
## 54              1 0.024479728 0.268950592 0.041832164
## 55              1 0.007067905 0.181168327 0.098101808
## 56              1 0.009946616 0.518576137 0.028288396
## 57              1 0.031627171 0.562782936 0.084788215
## 58              1 0.051863426 0.129156854 0.008223923
## 59              1 0.066200508 0.256367604 0.088645875
## 60              1 0.040683019 0.717935276 0.047193073
## 61              1 0.091287592 0.961409936 0.010910096
## 62              1 0.029360337 0.100140847 0.033327798
## 63              1 0.045906573 0.763222690 0.083741657
## 64              1 0.033239467 0.947966355 0.027684984
## 65              1 0.065087047 0.818634688 0.058703514
## 66              1 0.025801678 0.308292331 0.083673227
## 67              1 0.047854525 0.649579460 0.007115402
## 68              1 0.076631067 0.953355451 0.070277874
## 69              1 0.008424691 0.953732650 0.069882454
## 70              1 0.087532133 0.339979203 0.046396238
## 71              1 0.033907294 0.262474110 0.043693111
## 72              1 0.083944035 0.165453933 0.056217679
```

```
## 73            1 0.034668349 0.322168057 0.092848323
## 74            1 0.033377493 0.510125207 0.023046641
## 75            1 0.047635125 0.923968471 0.022181375
## 76            1 0.089219834 0.510959698 0.042021589
## 77            1 0.086433947 0.257621261 0.033352081
## 78            1 0.038998954 0.046460887 0.086480755
## 79            1 0.077732070 0.417856258 0.017719454
## 80            1 0.096061800 0.854001502 0.049331873
## 81            1 0.043465948 0.347230678 0.042971337
## 82            1 0.071251468 0.131442321 0.056426384
## 83            1 0.039999437 0.374486865 0.065616232
## 84            1 0.032535215 0.631420228 0.097855406
## 85            1 0.075708715 0.390078934 0.023216115
## 86            1 0.020269226 0.689627849 0.024081160
## 87            1 0.071112122 0.689413412 0.079683608
## 88            1 0.012169192 0.554900623 0.083167172
## 89            1 0.024548851 0.429624408 0.011350771
## 90            1 0.014330438 0.452720063 0.096331202
## 91            1 0.023962942 0.306443259 0.014732290
## 92            1 0.005893438 0.578353944 0.014362694
## 93            1 0.064228826 0.910370304 0.092522994
## 94            1 0.087626921 0.142604082 0.050703560
## 95            1 0.077891468 0.415047625 0.015485102
## 96            1 0.079730883 0.210925751 0.034830205
## 97            1 0.045527445 0.428750371 0.065982103
## 98            1 0.041008408 0.132689975 0.031177237
## 99            1 0.081087024 0.460096446 0.035157341
## 100           1 0.060493329 0.942957059 0.014784571
## 101           1 0.065472393 0.076197386 0.658877609
## 102           1 0.035319727 0.093290983 0.185069965
## 103           1 0.027026015 0.047067850 0.954378137
## 104           1 0.099268406 0.060358807 0.897848492
## 105           1 0.063349326 0.048498968 0.943697054
## 106           1 0.021320814 0.010880632 0.723690751
## 107           1 0.012937235 0.024772683 0.370357066
## 108           1 0.047811803 0.049851453 0.781017540
## 109           1 0.092407447 0.037286671 0.011149509
## 110           1 0.059876097 0.093469137 0.940308712
## 111           1 0.097617069 0.052398608 0.993749226
## 112           1 0.073179251 0.031714467 0.357405745
## 113           1 0.035672691 0.027796603 0.747635063
## 114           1 0.043147369 0.078754051 0.792909024
## 115           1 0.014821156 0.070246251 0.705859006
## 116           1 0.001307758 0.016502764 0.475825039
## 117           1 0.071556607 0.006445754 0.494654526
## 118           1 0.010318424 0.075470562 0.308052449
## 119           1 0.044628435 0.062041003 0.695012246
## 120           1 0.064010105 0.016957677 0.822793306
## 121           1 0.099183862 0.006221405 0.434717641
## 122           1 0.049559358 0.010902927 0.514732653
## 123           1 0.048434952 0.038171635 0.663010968
## 124           1 0.017344233 0.016931091 0.143166587
## 125           1 0.075482094 0.029865254 0.344487394
## 126           1 0.045389549 0.019220954 0.405763582
```

```
## 127           1 0.051116978 0.025717002 0.085311006
## 128           1 0.020754511 0.018123182 0.932571928
## 129           1 0.022865814 0.047731371 0.838384067
## 130           1 0.059571200 0.077073704 0.879433296
## 131           1 0.057487220 0.002778712 0.935712468
## 132           1 0.007706438 0.052731078 0.072460633
## 133           1 0.003554058 0.088031907 0.378759441
## 134           1 0.064279549 0.037306337 0.537864923
## 135           1 0.092861520 0.004795913 0.105050139
## 136           1 0.059809242 0.013862825 0.801687706
## 137           1 0.056090075 0.032149212 0.739641746
## 138           1 0.052602772 0.015483161 0.052149013
## 139           1 0.098509522 0.013222817 0.482169573
## 140           1 0.050764182 0.022130593 0.920517841
## 141           1 0.068278808 0.022638080 0.041528429
## 142           1 0.060154122 0.013141653 0.293991799
## 143           1 0.023886868 0.098156346 0.500850487
## 144           1 0.025816593 0.032701373 0.609748935
## 145           1 0.072930962 0.050693950 0.264249050
## 146           1 0.045257083 0.068144251 0.423098610
## 147           1 0.017512677 0.009916910 0.366563616
## 148           1 0.074669827 0.011890256 0.942505322
## 149           1 0.010498764 0.005043966 0.123723565
## 150           1 0.086454495 0.092925392 0.070032679
## attr(,"assign")
## [1] 0 1 2 3
```

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value Pr(>F)
## x1          1     0.0   0.014  0.0001 0.9935
## x2          1   177.4 177.373  0.8700 0.3525
## x3          1   200.1 200.093  0.9815 0.3235
## Residuals 146 29764.4 203.866
```

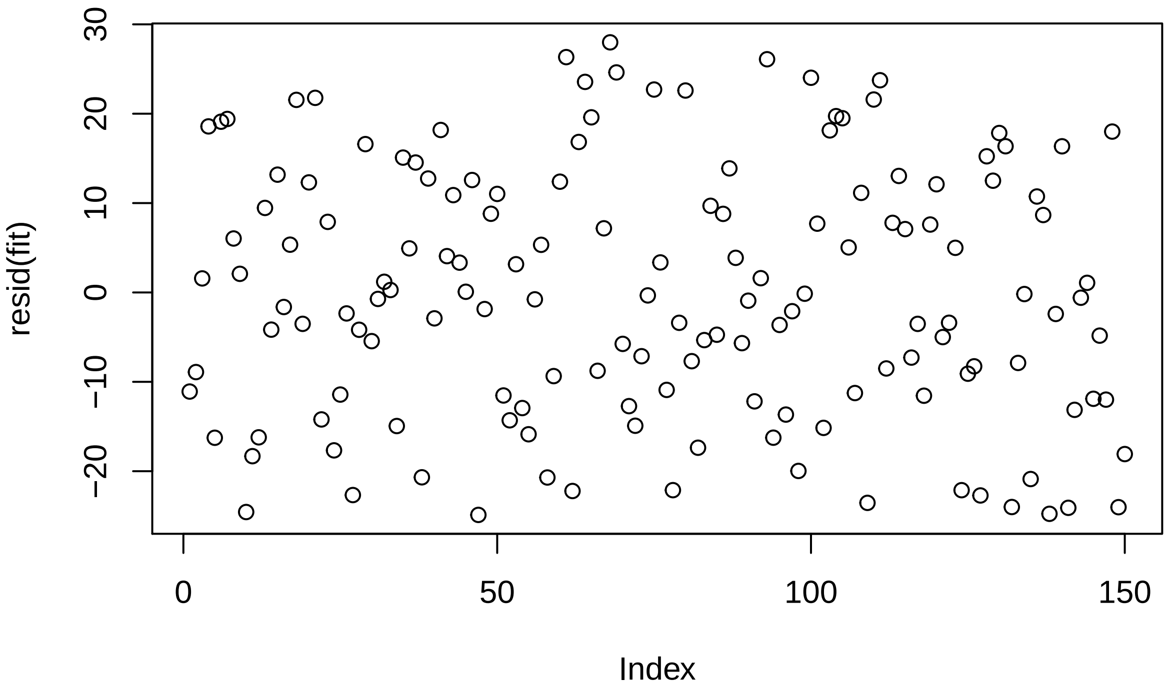
```r
cor(data)
```

```
##              x1          x2         x3            y
## x1  1.000000000  0.80292004 0.77986174 -0.000674168
## x2  0.802920043  1.00000000 0.79246795 -0.046267851
## x3  0.779861738  0.79246795 1.00000000  0.023727039
## y  -0.000674168 -0.04626785 0.02372704  1.000000000
```

The covariates seem more correlated with each other than with the outcome. Each has a correlation coefficient around 0.80 which points to multicollinearity. Based on the model, here are the relationships between the covariates and the outcome. For every unit increase in x1, there is an estimated 0.9003 unit increase in y holding x2 and x3 constant. This is not statistically significant at an alpha level of $\alpha = 0.10$ with a p-value of 0.770. For every unit increase in x2, there is an estimated 4.0704 unit decrease in y holding x1 and x3 constant. This is not statistically significant at an alpha level of $\alpha = 0.10$ with a p-value of 0.204. For every unit increase in x3, there is an estimated 2.9605 unit increase in y holding x1 and x2 constant. This is not statistically significant at an alpha level of $\alpha = 0.10$ with a p-value of 0.323. The model only explains 1.252% of the variance in y.

2.

```r
plot(resid(fit))
```



The residuals seem to have a random scatter, and no obvious pattern so there does not seem to be a problem

3.

```r
r <- resid(fit)
X <- fit$x
H <- X%*%solve(t(X)%*%X)%*%t(X)
h <- diag(H)
d <- r/(1-h)
d[7]
```

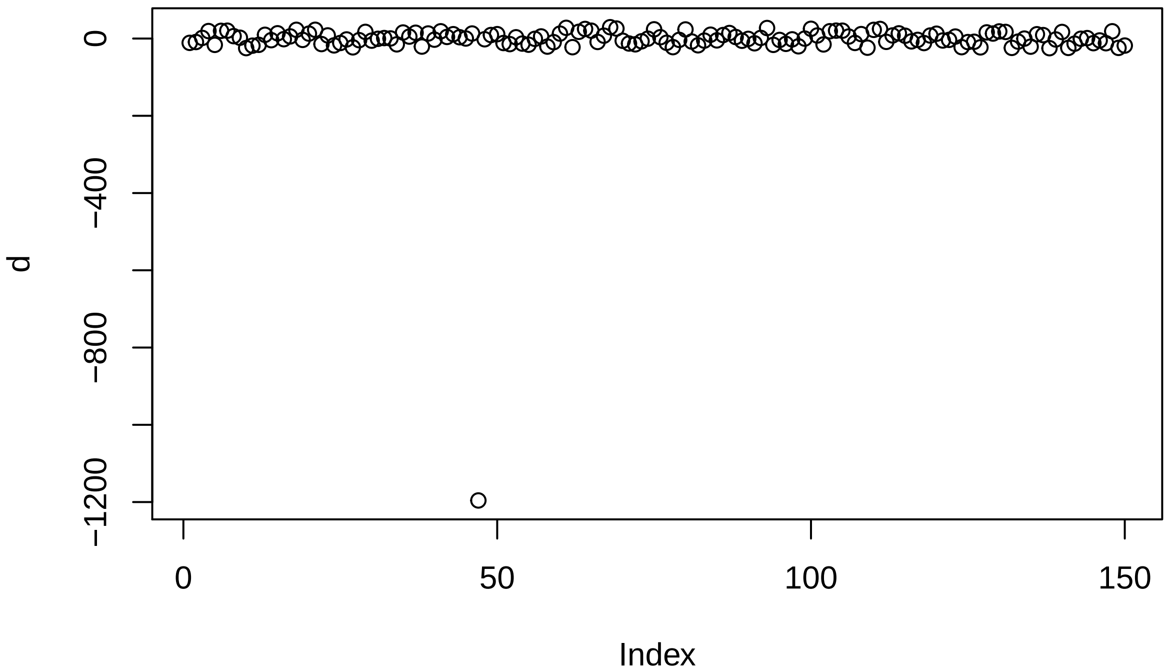
```
##       7
## 20.2805
```

```r
fit.without.7 <- lm(y~.,data=data[-7,])
data$y[7] - predict(fit.without.7,newdata=data[7,])
```

```
##       7
## 20.2805
```

4.

```r
plot(d)
```

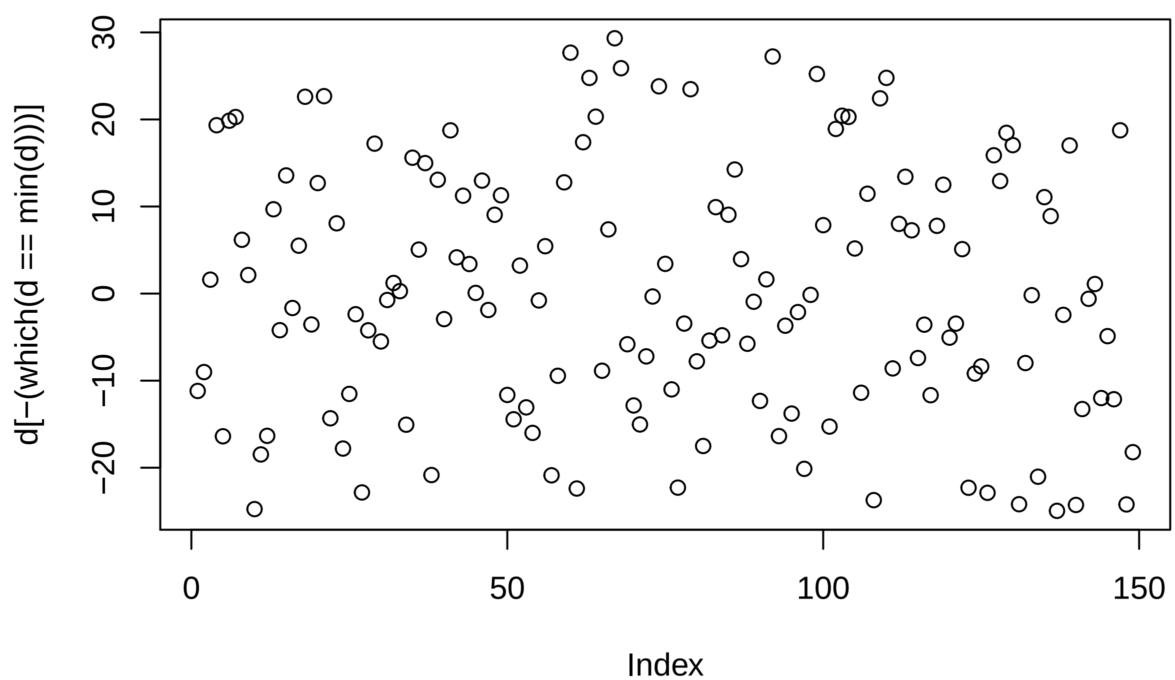


One point seems to have a much more extreme residual compared to the others, and so seems like an outlier. As a result, this seems to point to taking the one data point out. These residuals plotted are shown below, and seem to have a random scatter.

```
min(d)
```

```
## [1] -1195.831
```

```
plot(d[-(which(d == min(d)))])
```



As a result, it seems like we should be looking at a model without the outlier instead, so fit.without.7

```r
fit.without.47 <- lm(y~.,data=data[-47,])
summary(fit.without.47)
```

```
##
## Call:
## lm(formula = y ~ ., data = data[-47, ])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0259158 -0.0056705  0.0001048  0.0066726  0.0282432
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  0.002915   0.001980     1.472    0.143
## x1          49.993964   0.003567 14015.218   <2e-16 ***
## x2          49.998175   0.003844 13005.515   <2e-16 ***
## x3          49.994887   0.003434 14560.775   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00999 on 145 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 9.83e+07 on 3 and 145 DF,  p-value: < 2.2e-16
```

```r
anova(fit.without.47)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## x1          1  3718.9  3718.9  37263633 < 2.2e-16 ***
## x2          1  4551.6  4551.6  45607111 < 2.2e-16 ***
## x3          1 21159.2 21159.2 212016166 < 2.2e-16 ***
## Residuals 145     0.0     0.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cor(data[-47,])
```

```
##            x1         x2         x3         y
## x1  1.0000000 -0.3167408 -0.3300636 0.3554797
## x2 -0.3167408  1.0000000 -0.3328276 0.2604244
## x3 -0.3300636 -0.3328276  1.0000000 0.3997367
## y   0.3554797  0.2604244  0.3997367 1.0000000
```

Based on the new model, the correlation coefficients between each of the covariates is only around -0.32, so multicollinearity no longs seems to be a problem. The new model explains all of the variance in y: It has an R^2 of 1, and a residual sum of squares (RSS) of 0. For x1, x2, and x2, for every unit increase in one variable holding the other two constant, there is an estimated 49.99 unit increase in y.