T-Rex Analysis

Hartaig Singh

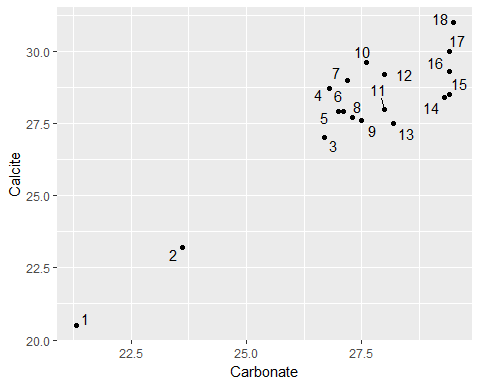
September 12, 2017

Import Data

Trex <- read.table("P8121F17-HW1-Trex.txt",header=T)

Scatterplot

ggplot(Trex,aes(Carbonate,Calcite,label=Id)) + geom\_point() + geom\_text\_repel(aes(label=Id))

 Points 1 and 2 appear furthest away from the rest of data

Linear Regression

m1 <- lm(Calcite~Carbonate,Trex)  
summary(m1)

##   
## Call:  
## lm(formula = Calcite ~ Carbonate, data = Trex)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.46796 -0.64104 -0.04927 0.67301 1.55856   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.4984 3.1766 -0.472 0.644   
## Carbonate 1.0703 0.1156 9.259 7.93e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9959 on 16 degrees of freedom  
## Multiple R-squared: 0.8427, Adjusted R-squared: 0.8329   
## F-statistic: 85.73 on 1 and 16 DF, p-value: 7.929e-08

**Interpretation**

*Carbonate:* For every unit increase in carbonate, calcite increases by 1.07 on average holding all other variables constant

*R^2:* 84.72% of the variation in calcite is explained by carbonate

Remove First Data point

m2 <- lm(Calcite~Carbonate,slice(Trex,2:18))  
summary(m2)

##   
## Call:  
## lm(formula = Calcite ~ Carbonate, data = slice(Trex, 2:18))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2799 -0.4816 -0.1364 0.7184 1.4871   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.6727 4.6247 0.578 0.572   
## Carbonate 0.9217 0.1663 5.541 5.65e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9807 on 15 degrees of freedom  
## Multiple R-squared: 0.6718, Adjusted R-squared: 0.6499   
## F-statistic: 30.7 on 1 and 15 DF, p-value: 5.653e-05

**Interpretation**

*Carbonate:* For every unit increase in Carbonate, Calcite increases by .9217 on average holding all other variables constant

*R^2:* 67.18% of the variation in Calcite is explained by Carbonate

Removing the first point decreased the R^2 by 20.7%. This is because removing points decreases the variance in carbonate giving you less of an ability to explain the variation in Calcite

Remove furthest two obersvations

m3 <- lm(Calcite~Carbonate,slice(Trex,3:18))  
summary(m3)

##   
## Call:  
## lm(formula = Calcite ~ Carbonate, data = slice(Trex, 3:18))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.1844 -0.7038 -0.1139 0.6854 1.5492   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.0589 6.1592 1.958 0.0705 .  
## Carbonate 0.5896 0.2196 2.684 0.0178 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8875 on 14 degrees of freedom  
## Multiple R-squared: 0.3398, Adjusted R-squared: 0.2926   
## F-statistic: 7.205 on 1 and 14 DF, p-value: 0.0178

**Interpretation**

*Carbonate:* For every unit increase in Carbonate, Calcite increases by .5896 on average holding all other variables constant

*R^2:* 33.98% of the variation in Calcite is explained by Carbonate

As expected, the R^2 decreases even more because we have increasingly less information to explain the variation in Calcite

Measures of Influence

influence.measures(m1)

## Influence measures of  
## lm(formula = Calcite ~ Carbonate, data = Trex) :  
##   
## dfb.1\_ dfb.Crbn dffit cov.r cook.d hat inf  
## 1 -1.333421 1.304917 -1.37520 2.126 9.17e-01 0.5578 \*  
## 2 -0.337365 0.325423 -0.36886 1.439 7.06e-02 0.2507 \*  
## 3 -0.008034 0.006640 -0.02023 1.212 2.18e-04 0.0623   
## 4 0.149131 -0.119777 0.41915 0.867 7.93e-02 0.0605   
## 5 0.033535 -0.024527 0.12524 1.167 8.22e-03 0.0578   
## 6 0.021533 -0.014467 0.09723 1.181 4.99e-03 0.0568   
## 7 0.063048 -0.036501 0.36244 0.916 6.11e-02 0.0561   
## 8 -0.000621 0.000257 -0.00495 1.205 1.31e-05 0.0557   
## 9 -0.002241 -0.003788 -0.08154 1.187 3.52e-03 0.0557   
## 10 -0.008926 0.039580 0.41525 0.846 7.71e-02 0.0561   
## 11 0.025153 -0.033770 -0.12021 1.175 7.59e-03 0.0603   
## 12 -0.039551 0.053101 0.18902 1.126 1.84e-02 0.0603   
## 13 0.096300 -0.119128 -0.32698 0.997 5.16e-02 0.0641   
## 14 0.347068 -0.378048 -0.55420 0.917 1.39e-01 0.1039   
## 15 0.371414 -0.402824 -0.57489 0.917 1.49e-01 0.1091   
## 16 0.158105 -0.171476 -0.24472 1.198 3.09e-02 0.1091   
## 17 -0.007464 0.008095 0.01155 1.277 7.12e-05 0.1091   
## 18 -0.235861 0.254815 0.35492 1.133 6.31e-02 0.1147

Observtations 1 and 2 are flagged by R as influential points as expected

Remove first observation measure of influence

influence.measures(m2)

## Influence measures of  
## lm(formula = Calcite ~ Carbonate, data = slice(Trex, 2:18)) :  
##   
## dfb.1\_ dfb.Crbn dffit cov.r cook.d hat inf  
## 1 -2.2426 2.2066 -2.3330 1.517 2.228990 0.5578 \*  
## 2 -0.0594 0.0556 -0.0932 1.248 0.004622 0.0914   
## 3 0.2695 -0.2507 0.4483 0.944 0.093364 0.0856   
## 4 0.0517 -0.0472 0.1001 1.220 0.005326 0.0756   
## 5 0.0330 -0.0297 0.0705 1.225 0.002653 0.0715   
## 6 0.1528 -0.1353 0.3684 0.960 0.064179 0.0680   
## 7 -0.0131 0.0113 -0.0367 1.224 0.000719 0.0650   
## 8 -0.0255 0.0199 -0.1096 1.191 0.006347 0.0608   
## 9 0.0687 -0.0476 0.4157 0.855 0.077486 0.0596   
## 10 0.0139 -0.0203 -0.1252 1.180 0.008251 0.0604   
## 11 -0.0210 0.0306 0.1888 1.131 0.018362 0.0604   
## 12 0.0793 -0.0956 -0.3281 0.992 0.051877 0.0643   
## 13 0.3832 -0.4030 -0.5508 0.995 0.141376 0.1266   
## 14 0.4111 -0.4311 -0.5727 1.006 0.152889 0.1358   
## 15 0.1436 -0.1506 -0.2000 1.281 0.021053 0.1358   
## 16 -0.0689 0.0722 0.0959 1.317 0.004910 0.1358   
## 17 -0.3895 0.4073 0.5277 1.077 0.133580 0.1455

The second obersvation which we guessed to be influential is flagged and rightfully so since it has a relatively high hat value indicating it being far from the rest of the data, high dffit, and dfbetas showing it having a heavy impact on the estimates and fitting line.

Remove first 2 observation measure of influence

influence.measures(m3)

## Influence measures of  
## lm(formula = Calcite ~ Carbonate, data = slice(Trex, 3:18)) :  
##   
## dfb.1\_ dfb.Crbn dffit cov.r cook.d hat inf  
## 1 -0.3654 0.35583 -0.4475 1.209 0.100273 0.1700   
## 2 0.3504 -0.34048 0.4413 1.171 0.096924 0.1544   
## 3 -0.0251 0.02428 -0.0341 1.327 0.000625 0.1268   
## 4 -0.0397 0.03820 -0.0566 1.305 0.001719 0.1149   
## 5 0.2440 -0.23387 0.3697 1.089 0.067508 0.1042   
## 6 -0.1036 0.09867 -0.1692 1.229 0.015103 0.0947   
## 7 -0.1125 0.10529 -0.2283 1.150 0.026824 0.0794   
## 8 0.1849 -0.17046 0.4396 0.888 0.087647 0.0736   
## 9 -0.0101 0.00412 -0.1667 1.162 0.014496 0.0625   
## 10 0.0114 -0.00463 0.1871 1.143 0.018126 0.0625   
## 11 0.0507 -0.06403 -0.3752 0.925 0.065481 0.0644   
## 12 0.3890 -0.40072 -0.5113 1.136 0.127496 0.1620   
## 13 0.4084 -0.41982 -0.5210 1.174 0.133307 0.1783   
## 14 0.0402 -0.04133 -0.0513 1.409 0.001415 0.1783   
## 15 -0.2716 0.27914 0.3464 1.299 0.061971 0.1783   
## 16 -0.8711 0.89374 1.0833 0.767 0.460917 0.1957

There are no influential points since it is clear from the scatter plot that every other point aside from the first two are clustered together

rstandard(m1) - rstudent(m1)

## 1 2 3 4 5   
## 0.0186796021 -0.0121630132 -0.0025582835 -0.0825848115 0.0121892124   
## 6 7 8 9 10   
## 0.0108687893 -0.0531683493 -0.0006677216 -0.0097170920 -0.0931070645   
## 11 12 13 14 15   
## -0.0119240579 0.0105574933 0.0213953116 0.0778908762 0.0807567479   
## 16 17 18   
## -0.0114435601 0.0010812400 0.0008409477

The more observations that are deleted, the more emphasis the measures of influence put on the next observation which makes sense since these metrics are all relative to the data set. This tends to be an issue with small data sets where any slight deviation from a group will trigger potential red flags causing spurious influences.

The deleted studentizes residuals do not really tell us much in terms of outliers since the values above are all still relatively small.