2/2/2010 36-402/608 ADA-II H. Seltman

Breakout #7 Results

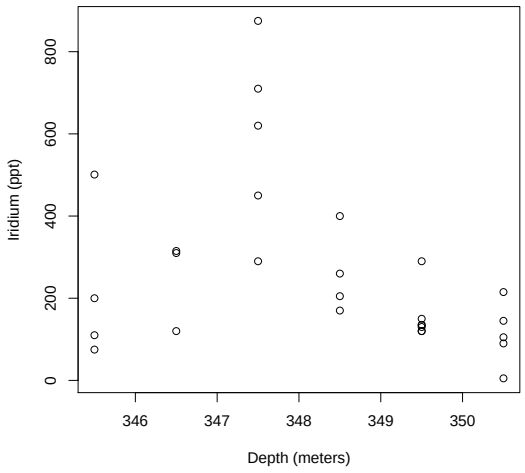
Geologists know that younger rocks are generally deposited on top of older rocks, so depth can, with appropriate controls, be used as a surrogate for age. Another geological fact is that the element iridium is quite rare in the Earth's crust, except for that coming from certain volcanic eruptions. Another source of iridium is the impact of comets or meteors.

The data from this study were collected in an attempt to elucidate the nature of the high iridium levels found at the so-called K-T boundary, which is the zone below which are rocks of the Cretaceous period, and above which are rocks of the Tertiary period. It is now generally believed that a large impact caused the deposit of iridium and the extinction of the dinosaurs at that time, partly because the "spike" of iridium is narrow in time.

The data were collected at 6 depths (coded A to F and corresponding to 6 consecutive time periods). For each sample the depth, and iridium level (in parts per trillion) were recorded.

```
summary(extinct)
# depth    iridium
# A:4   Min.   :  5.0
# B:3   1st Qu.:120.0
# C:5   Median :185.0
# D:4   Mean   :259.0
# E:7   3rd Qu.:311.2
# F:5   Max.   :875.0
```

```
plot(c(345.5,346.5,347.5,348.5,349.5,350.5)[as.numeric(depth)],
     iridium, xlab="Depth (meters)", ylab="Iridium (ppt)")
```



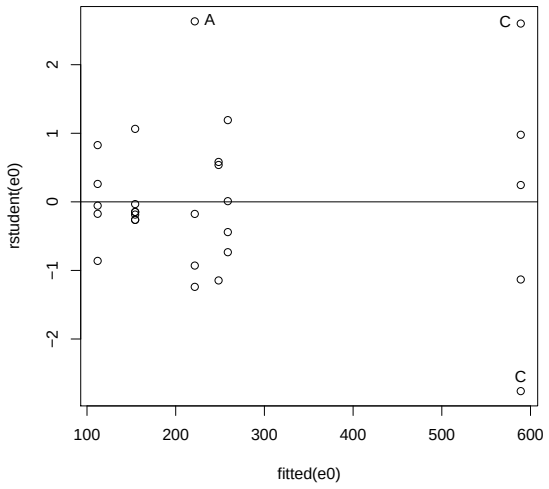
The question of interest to geologists is whether there is a real rise in iridium at some depth.

Question 1: Why is linear regression not appropriate here, even with adding a depth^2 term?

In R, `aov()`, not `anova()`, is used to perform a standard Analysis of Variance:

```
e0 = aov(iridium ~ depth, extinct)
summary(e0)
#             Df Sum Sq Mean Sq F value    Pr(>F)
#depth         5 735267  147053  7.7058 0.0002568 ***
#Residuals    22 419834   19083

plot(fitted(e0), rstudent(e0));abline(h=0)
identify(fitted(e0), rstudent(e0), )
```



Question 2: How many vertical bands are seen in the residual plot and why? What are the meanings of all of the numbers in the ANOVA table?

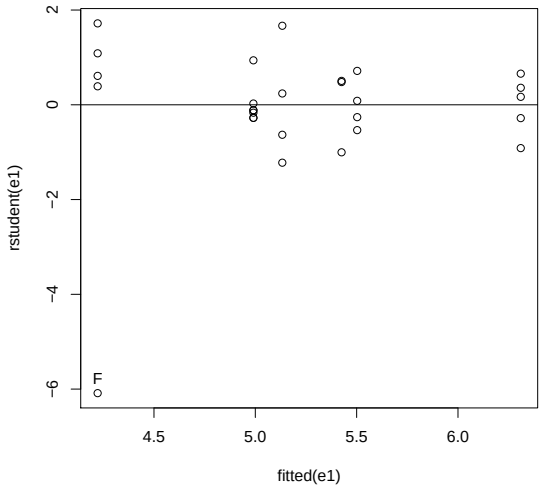
```
summary(influence.measures(e0))
# Potentially influential observations of
#        aov(formula = iridium ~ depth, data = extinct) :
#
#   dfb.1_  dfb.dptB dfb.dptC dfb.dptD dfb.dptE dfb.dptF
# 4  1.52_* -0.99    -1.13_*  -1.07_*  -1.21_*  -1.13_*
# 6  0.00    0.29     0.00     0.00     0.00     0.00
#   dffit cov.r   cook.dhat
# 4  1.52  0.32    0.30  0.25
# 6  0.38  1.83_*  0.02  0.33

extinct[apply(influence.measures(e0)$is.inf,1,any,na.rm=TRUE),]
#   depth iridium
# 4     A     501
# 6     B     310
```

It is generally recommended to try a log transformation when the ratio of the largest to the smallest values for a variable is at least, say, 20.

```
e1 = aov(log(iridium) ~ depth, extinct)
summary(e1)
#             Df  Sum Sq Mean Sq F value  Pr(>F)
# depth        5 11.7587  2.3517  3.8695 0.01146 *
# Residuals   22 13.3709  0.6078

plot(fitted(e1), rstudent(e1));abline(h=0)
identify(fitted(e1), rstudent(e1), depth)
```



Question 4: Which form of "iridium" do you prefer and why?

```
tmpm = with(extinct, aggregate(iridium, list(depth), mean))$x
tmpm # 221.5000 248.3333 589.0000 258.7500 154.2857 112.0000

tmpn = with(extinct, aggregate(iridium, list(depth), length))$x
tmpn # 4 3 5 4 7 5
```

Our contrast hypothesis is "Is depth 3 iridium different from the other times?"

Question 5: How would you calculate, G, the estimate of the difference in mean iridium for time 3 vs. the others? What is the "C" vector?

Question 5: How would you calculate the variance of G?

Question 6: How would you calculate the T statistic and p-value?