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

Breakout #7 Results

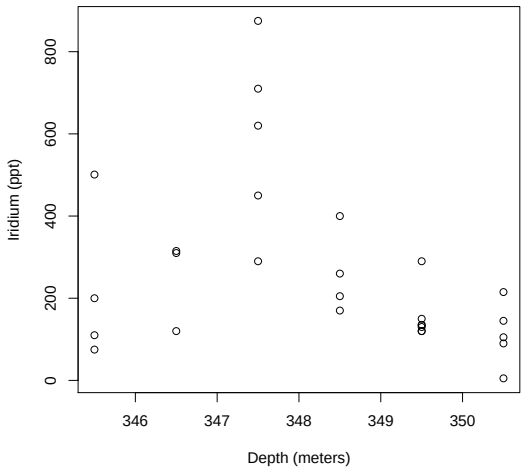
Geologists know that younger rocks are generally deposited on top of older rocks, so depth can, with appropriate controls, be used as a surrogate for age. Another geological fact is that the element iridium is quite rare in the Earth's crust, except for that coming from certain volcanic eruptions. Another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

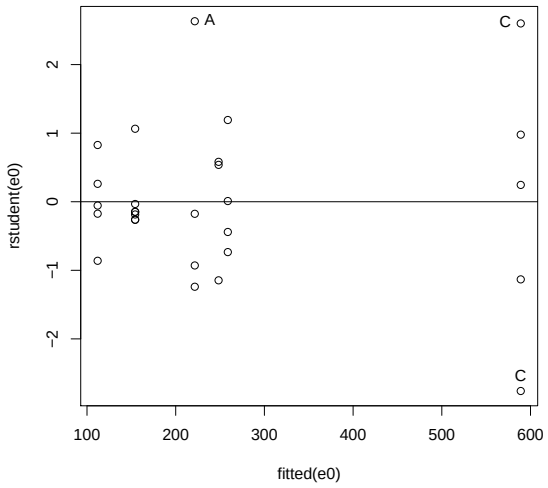
Question 1: Why is linear regression not appropriate here, even with adding a depth^2 term?

Depth is inherantly quantitative so you might think regression would make the most sense, but scientifically, we don't expect a linear or quadratic change in iridium over time because we expect that it came from the meteor impact catastrophe which occured over a very small time period. So we expect one time period to have a much higher iridium concentration than the rest and need to treat each time period as if it had a separate mean iridium level.

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?

The x-axis is fitted (predicted) iridium. In the ANOVA model the best prediction for an observation from any group is the mean of the that group (and we attribute deviations from the group mean to "error" of all sorts). So there are 6 possible values of fitted values for the 6 group means.

Under "within group deviations", df is $(n_i - 1)$ from each of the six groups, i.e., $\sum_{i=1}^{I}(n_i - 1) = 22$, SS is the sum of squared deviations of observations from their group means, and MS = SS/df. Under "between group deviations", df is $I - 1 = 5$ because the six deviations of group means from the grand mean are constrained to add to 0 so only five are free to vary, SS is the sum of squares of these deviations, and MS = SS/df.

The F statistic equals MWB/MSW, and under the null hypothesis that all six group population means are equal, the numerator and denominator are both independent estimators of the common variance, σ^2. Under H_0 this F statistic (ratio) follows an F distribution with 5 and 22 df. The p-value tells us that an F value of 7.7058 or greater would occur on 0.00002568 of the time (over repeated experiments) if the null hypothesis were true. So we conclude that the null hypothesis is probably false, and at least one depth has a different population mean of iridium than the other depths.

```
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
```

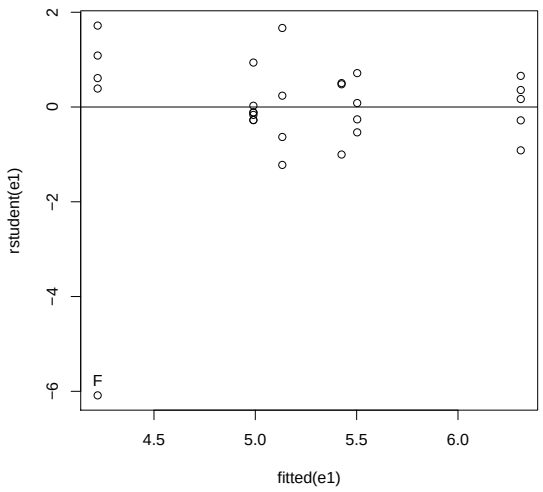
These two observations have neither high leverage (hat is not flagged) nor high overall influence (Cook's D is not flagged), so we probably don't need to be too concerned. If we are very careful, we would compare the results with and without this data points and ask the subject matter expert whether the conclusions differ in a meaningful way. If not we will ignore the "outliers". If so, we may need to investigate the cause of the outliers or report results with and without them.

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?

The log form of iridium achieves equal variance (we should also check the quantile normal plot to see if Normality improves or gets worse), but at the expense of one extreme outlier. If we have reason to drop the outlier, the log is best. If not we will need to "invoke" the robustness of ANOVA to unequal variance. Roughly, variance ratios between groups of less than 3 have minimal effect on standard errors, confidence limits and p-values, and, while larger ratios do affect these quantities, at least for p-values, changing alpha to alpha/2 protects our type-1 error in all but the most extreme cases. So we could say that with $p << 0.025$ we can safely reject H_0 for the non-logged data.

```
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?

We should probably use the logged data, but here is what you get for the non-logged data.

We can test $\mu_3 = \frac{\mu_1+\mu_2+\mu_4+\mu_5+\mu_6}{5}$ by choosing $C = (-\frac{1}{5}, -\frac{1}{5}, +1, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5},)$ so that $\gamma = \sum_{i=1}^{6} C_i\mu_i = 0$ reflects our null hypothesis.

We can estimate gamma with $G = \sum_{i=1}^{6} C_i\bar{Y}_i = 0$ which is

```
sum(tmpm * c(-1/5,-1/5, 1, -1/5,-1/5,-1/5)) # [1] 390.0262
```

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

We estimate σ^2 with s_p^2=MSW=19083 with 22 df.

$$\text{Var}(G) = \sum_{i=1}^{6} C_i^2 \frac{s_p^2}{n_i} = 19083[(-1/5)^2/4 + \cdots + (-1/5)^2/5] = 4714.41$$

The SE(G) = $\sqrt{4714.41} = 68.66$.

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

Using the general definition of t-statistic, $T = \frac{G-\gamma}{\text{SE}(G)} \sim t_{df}$.

Under $H_0 : \gamma = 0$, we have t=390.0262/68.66=5.68.

Clearly $p << 0.05$. To get an exact p-value, use

```
2 * pt(-5.68, 22)
```

to find p= 0.000010333 and reject H_0.