Biostatistics 140.754
Advanced Methods in Biostatistics IV

Jeffrey Leek

Assistant Professor
Department of Biostatistics
jleek@jhsph.edu

Lecture 8

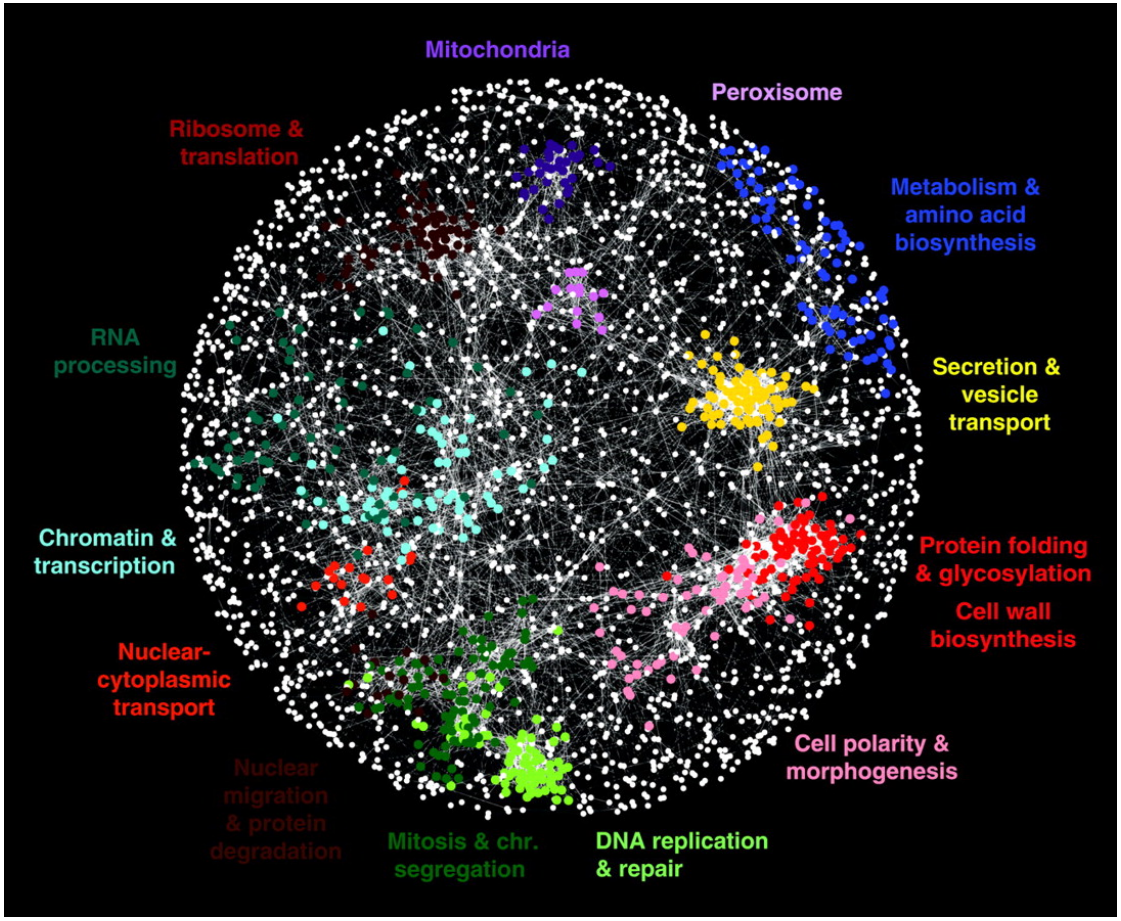
Tip + Paper

Tip In data analysis - particularly for complex high-dimensional data - it is frequently better to choose simple models for clearly defined parameters. With a lot of data, there is a strong temptation to go overboard with statistically complicated models; the danger of over-fitting/over-interpreting is extreme. The most reproducible results are often produced by sensible and statistically "simple" analyses (Note: being sensible and simple does not always lead to higher profile results).

Beware ridiculograms! http://www.youtube.com/watch?v=YS-asmU3p_4&feature=channel_video_title

Paper of the Day:
"The genetic landscape of a cell"
http://www.sciencemag.org/content/327/5964/425.short

A ridiculogram! (Apologies to C. Myers)



Outline For Today

- Data analysis

The Earthquake Data

The earthquake data available from the class website look like this:

```
Src,Eqid,Version,Datetime,Lat,Lon,Magnitude,Depth,NST,Region
ak,10208729,1,"Monday, April 11, 2011 21:28:47 UTC",59.8229,-150.2962,2.6,23.80,23,"Kenai
Peninsula, Alaska"
us,c0002nwt,6,"Monday, April 11, 2011 21:21:19 UTC",38.8311,141.9716,4.7,58.00,39,"near the east
coast of Honshu, Japan"
ak,10208720,1,"Monday, April 11, 2011 21:19:21 UTC",61.1910,-150.8566,1.5,35.00,13,"Southern
Alaska"
us,c0002nvr,4,"Monday, April 11, 2011 21:03:02 UTC",36.3607,145.7883,4.7,15.10,19,"off the east
coast of Honshu, Japan"
```

Read the data into R

```
> edata = read.csv("eqs7day-M1.txt",header=T)
> dim(edata)

[1] 1007 10
```

Also, the dates are for April 11. Did any of you re-download the data from the website? http://www.data.gov/raw/34.

Earthquake depth/magnitude

The question of interest is: what is the relationship between earthquake magnitude and earthquake depth?

What are earthquake depth/magnitude? Let's start here: http://earthquake.usgs.gov/earthquakes/glossary.php

- The **focus** of an earthquake is the place where the energy is initially released (i.e., where the rocks first break).
- The **depth** where the earthquake begins to rupture in km. This depth may be relative to mean sea-level or the average elevation of the seismic stations which provided arrival-time data for the earthquake location. "Earthquakes can occur anywhere between the Earth's surface and about 700 kilometers below the surface." [1]
- Earthquake **magnitude** is a logarithmic measure of earthquake size. In simple terms, this means that at the same distance from the earthquake, the shaking will be 10 times as large during a magnitude 5 earthquake as during a magnitude 4 earthquake.

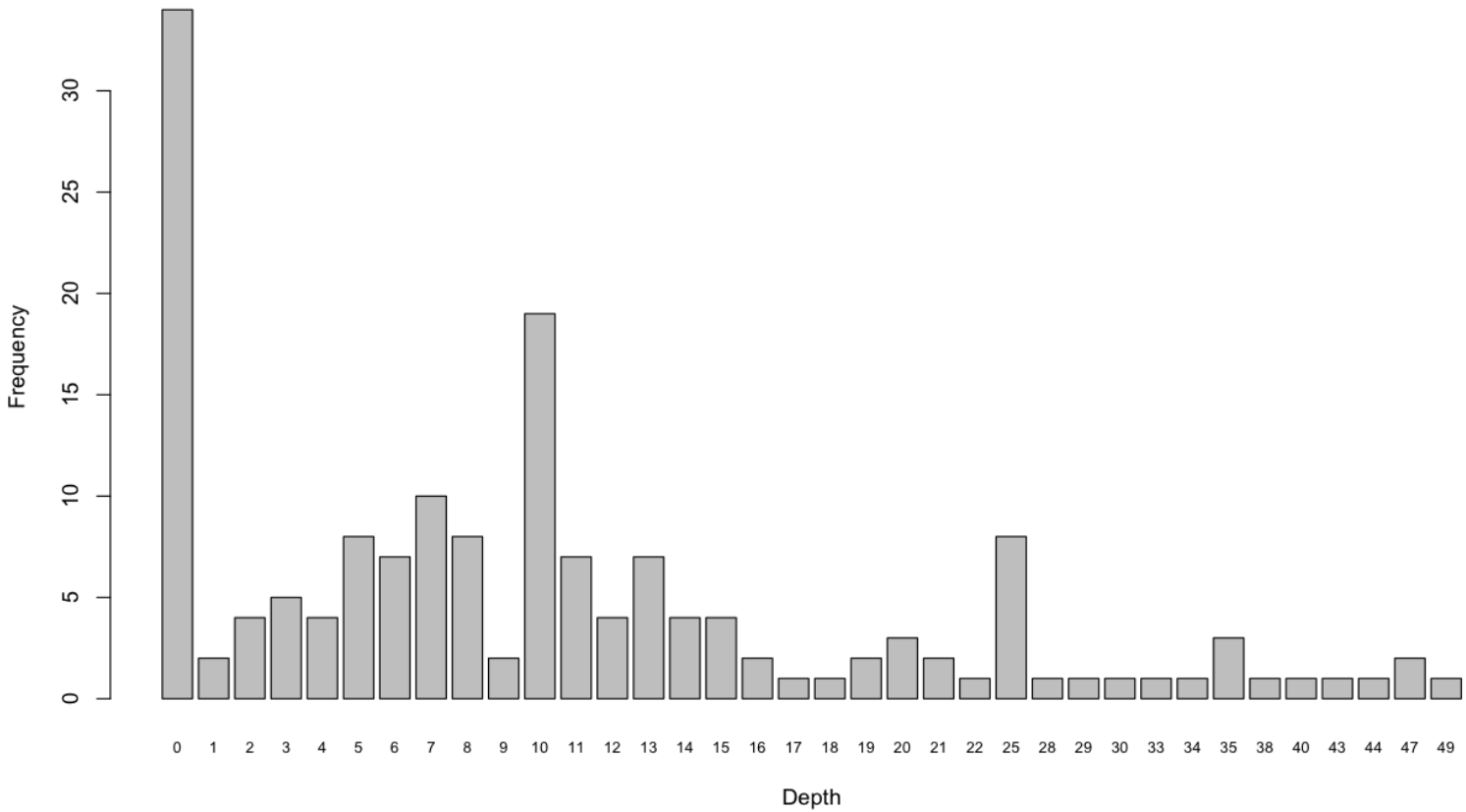
[1] http://earthquake.usgs.gov/learn/topics/seismology/determining_depth.php

Depth Values

"Sometimes when depth is poorly constrained by available seismic data, the location program will set the depth at a fixed value. For example, 33 km is often used as a default depth for earthquakes determined to be shallow, but whose depth is not satisfactorily determined by the data, whereas default depths of 5 or 10 km are often used in mid-continental areas and on mid-ocean ridges since earthquakes in these areas are usually shallower than 33 km."

Depth Values

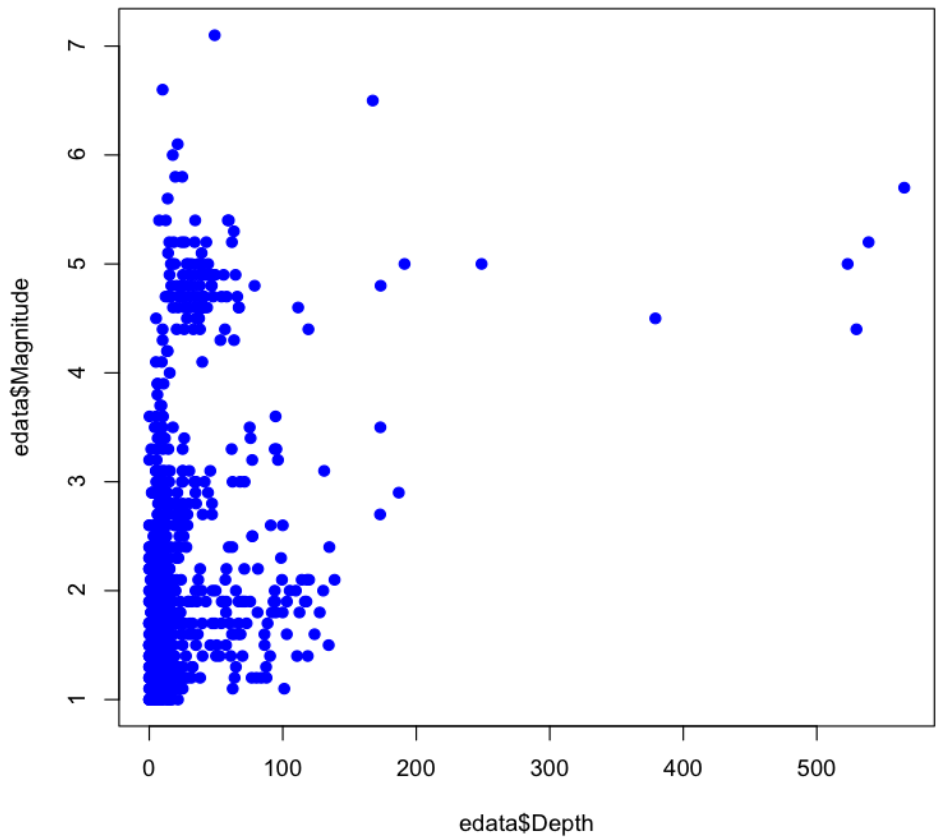
```
> summary(edata$Depth)
Min.  1st Qu.  Median Mean 3rd Qu.  Max.
0.00 4.70 9.90 22.54 22.50 565.50

> integerdepths = table(edata$Depth[which(edata$Depth==round(edata$Depth))])
> barplot(integerdepths[as.numeric(names(integerdepths)) <
50],cex.names=0.7,ylab="Frequency",xlab="Depth")
```



Magnitude Values and Depth vs. Magnitude Plot

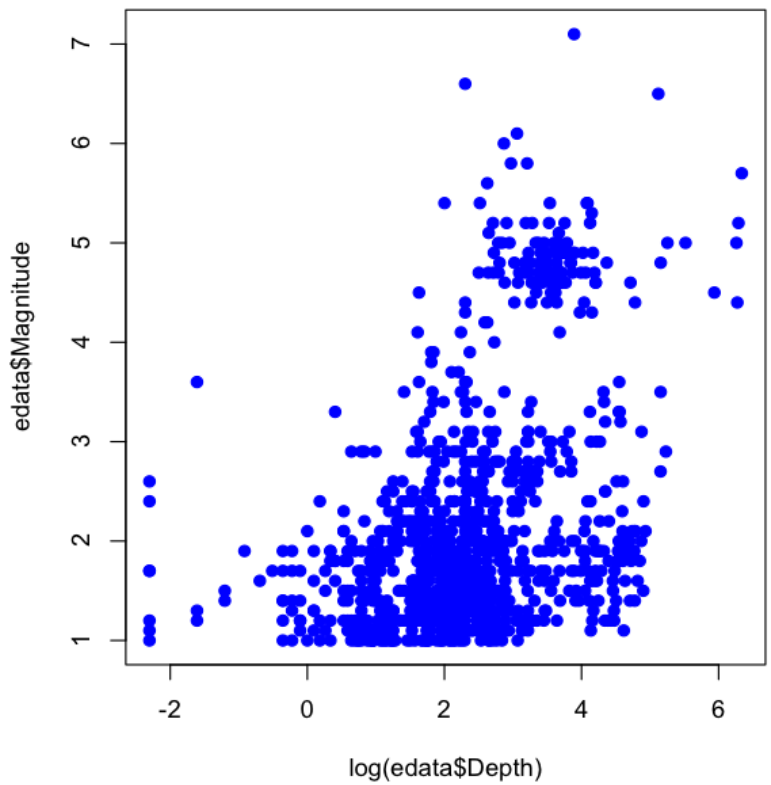
```
> summary(edata$Magnitude)
Min.  1st Qu.  Median Mean 3rd Qu.  Max.
1.000 1.400 1.800 2.189 2.600 7.100
```

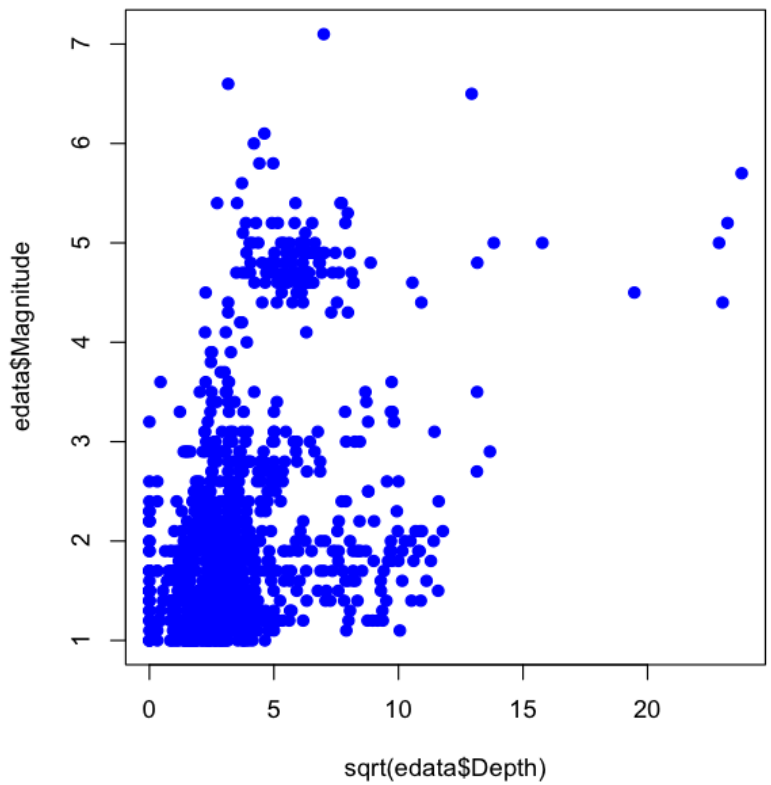
```
> plot(edata$Depth,edata$Magnitude,pch=19,col="blue")
```



f(Depth) versus Magnitude

```
> par(mfrow=c(1,2))
> plot(log(edata$Depth),edata$Magnitude,pch=19,col="blue")

> plot(sqrt(edata$Depth),edata$Magnitude,pch=19,col="blue")
```





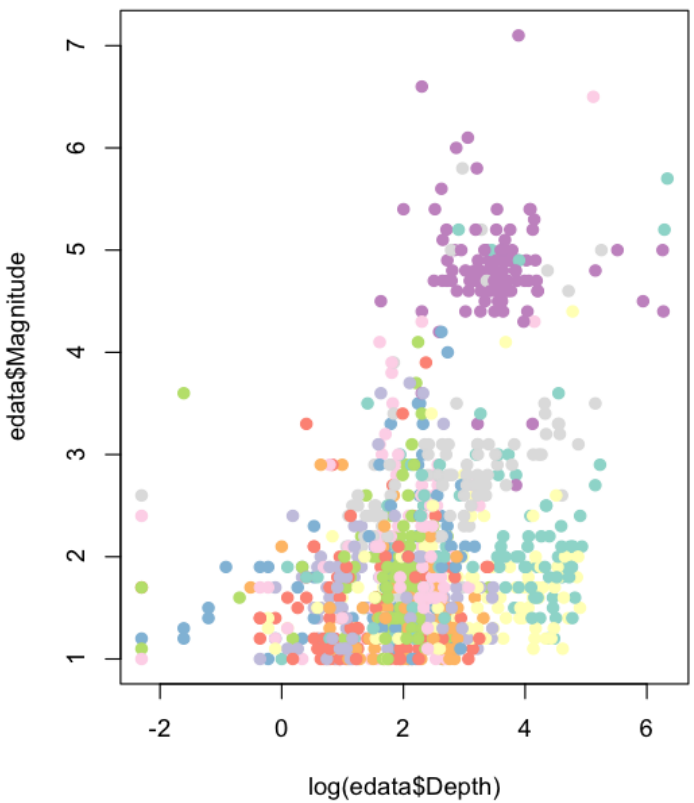
The other variables

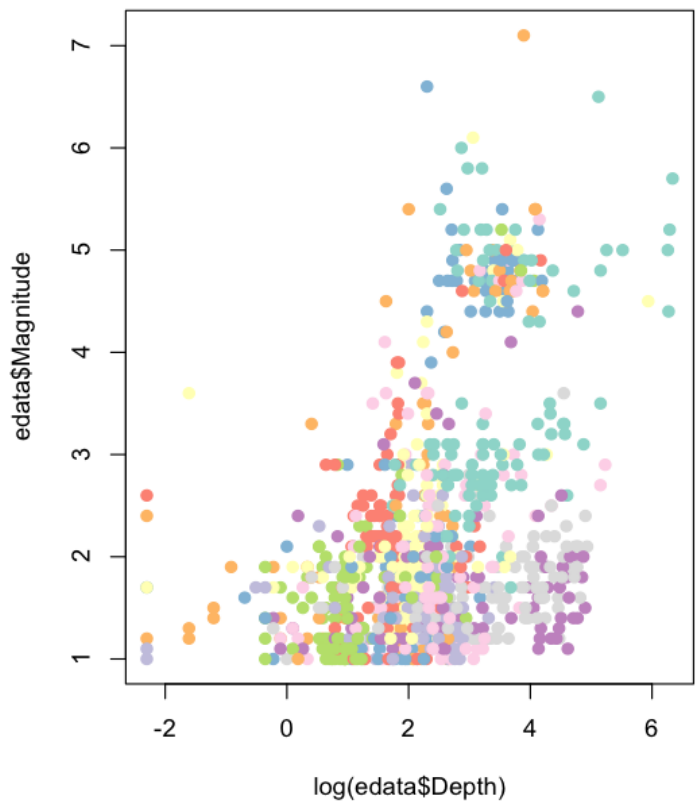
```
> names(edata)
[1] "Src" "Eqid" "Version"
[4] "Datetime" "Lat" "Lon"
[7] "Magnitude" "Depth" "NST"
[10] "Region"
```

- Src - the source of the information presented in the table
- Eqid - Event/Earthquake ID - a unique identifier for each earthquake
- Version - the version number as the data are updated - higher numbers mean more updates
- Datetime - Time of the initial rupture in UTC
- Lat - the latitude of the epicenter of the earthquake (positive is N, negative is S)
- Lon - the longitude of the epicenter of the earthquake from Greenwich (positive is E, negative is W)
- NST - Number of seismic stations which reported P- and S-arrival times for this earthquake
- Region - The region name is an automatically generated name from the Flinn-Engdahl (F-E) seismic and geographical regionalization scheme, proposed in 1965, defined in 1974 and revised in 1995.

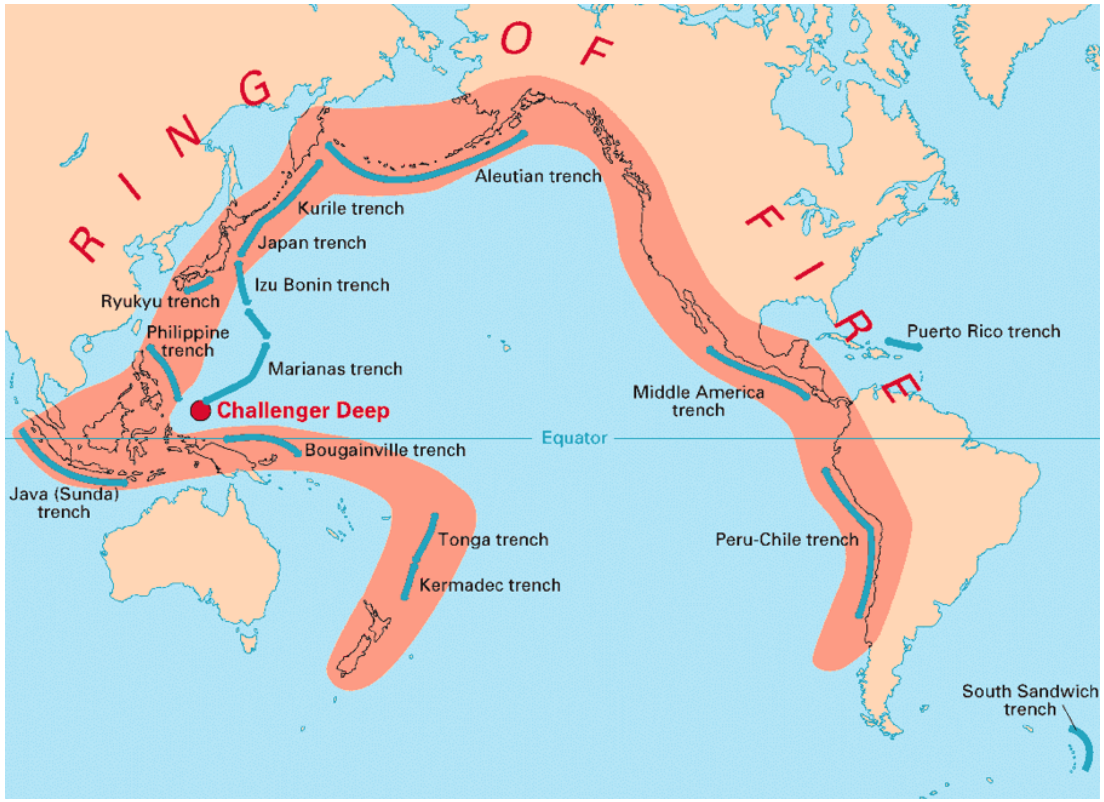
log(Depth) versus Magnitude - colored by Latitude, Longitude

```
> library(Hmisc)
> library(RColorBrewer)
> cols1 = c(brewer.pal(12,"Set3"),brewer.pal(6,"Dark2"))
> lonFactor = cut2(edata$Lon,g=10)
> latFactor = cut2(edata$Lat,g=10)
> par(mfrow=c(1,2))
> plot(log(edata$Depth),edata$Magnitude,pch=19,col=cols1[as.numeric(lonFactor)])
> plot(log(edata$Depth),edata$Magnitude,pch=19,col=cols1[as.numeric(latFactor)])
```





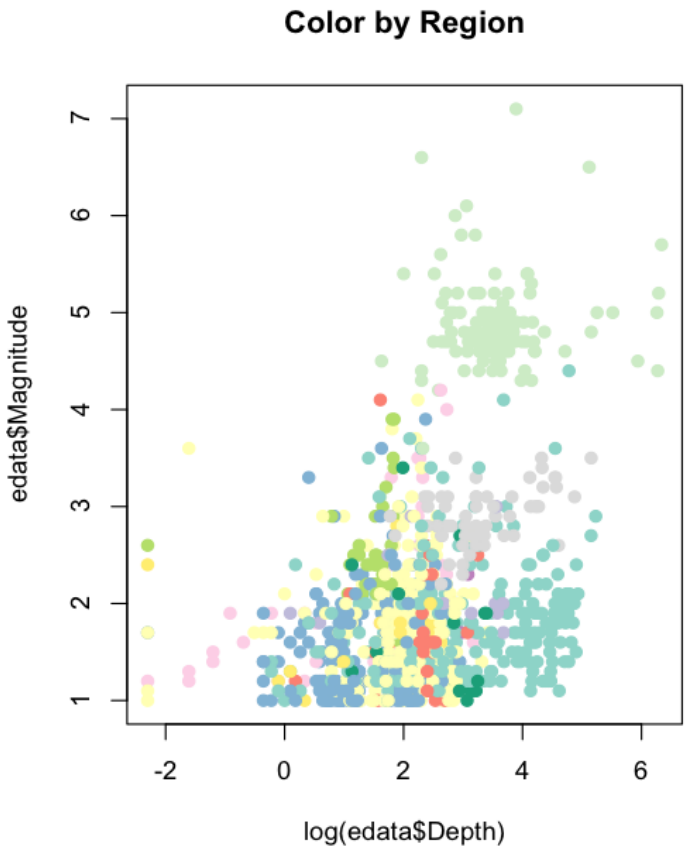
The Ring of Fire

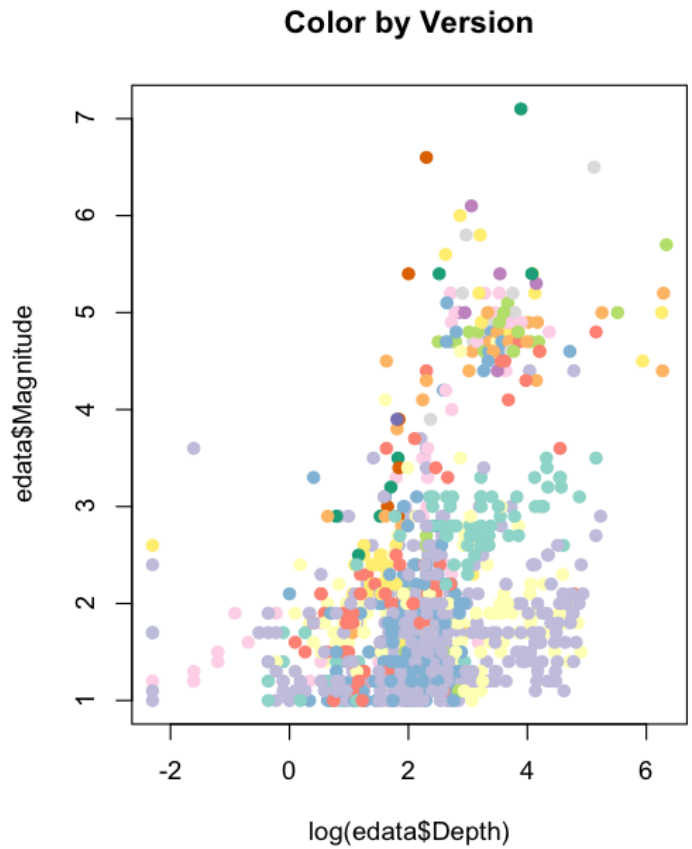


also...

log(Depth) versus Magnitude - colored by Region/Version

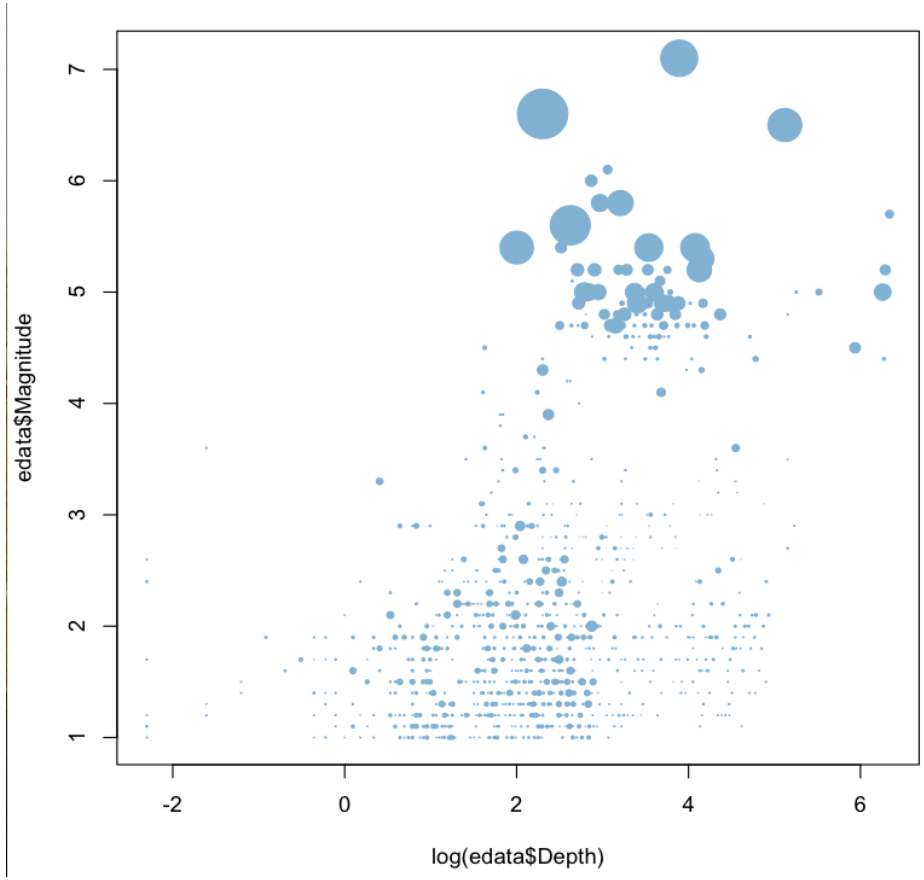
```
> library(RColorBrewer)
> cols1 = c(brewer.pal(12,"Set3"),brewer.pal(6,"Dark2"))
> par(mfrow=c(1,2))
> plot(log(edata$Depth),edata$Magnitude,pch=19,col=cols1[as.numeric(as.factor(edata$Src))])
> plot(log(edata$Depth),edata$Magnitude,pch=19,col=cols1[as.numeric(as.factor(edata$Version))])
```





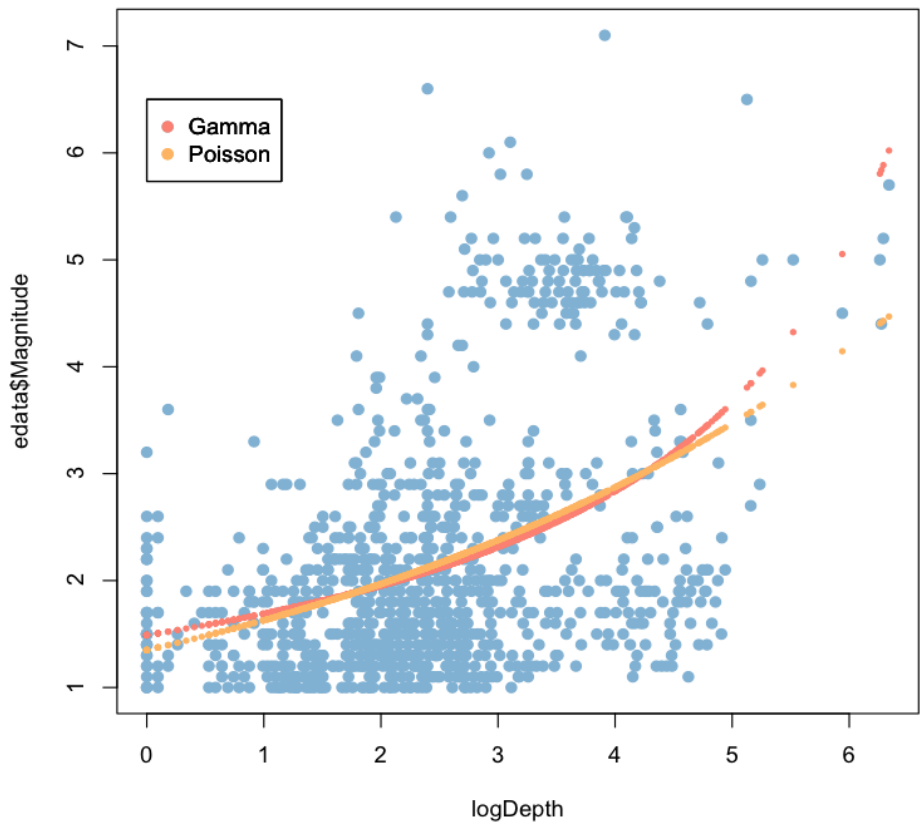
log(Depth) versus Magnitude - sized by the Number of Stations

```
> plot(log(edata$Depth),edata$Magnitude,pch=19,col=cols1[5],cex=5*edata$NST/max(edata$NST))
```



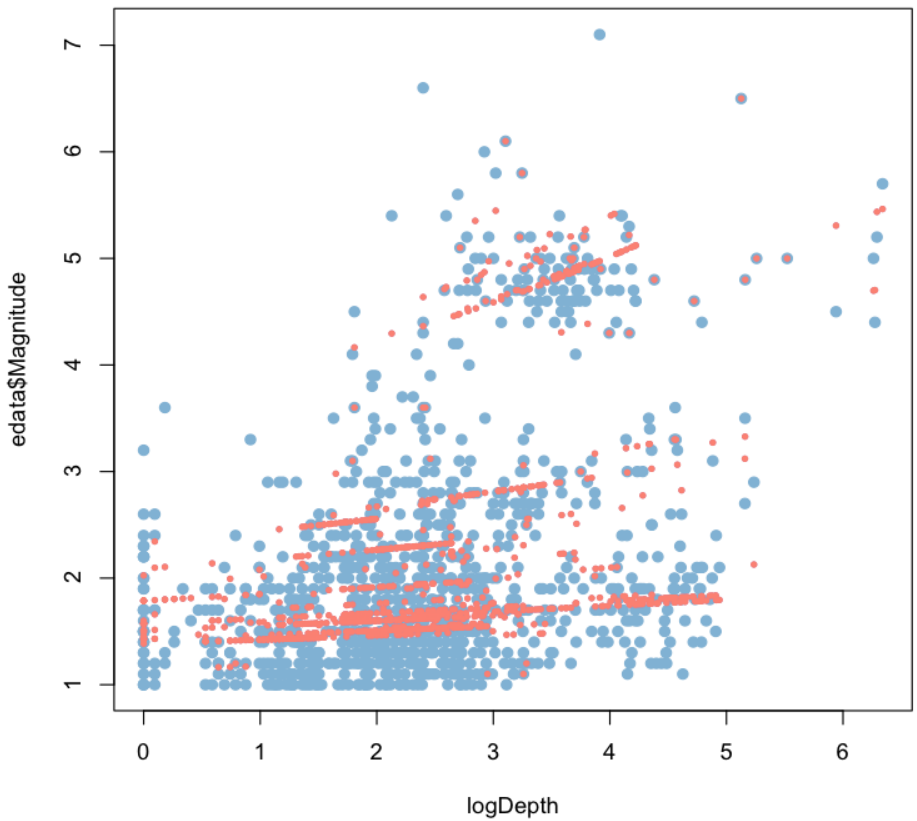
Model Magnitude and log(Depth)

```
> logDepth <- log(edata$Depth + 1)
> glm1 <- glm(edata$Magnitude ~ logDepth,family="Gamma")
> glm2 <- glm(edata$Magnitude ~ logDepth,family="poisson")
> plot(logDepth,edata$Magnitude,pch=19,col=cols1[5])
> points(logDepth,glm1$fitted,pch=19,cex=0.5,col=cols1[4])
> points(logDepth,glm2$fitted,pch=19,cex=0.5,col=cols1[6])
```



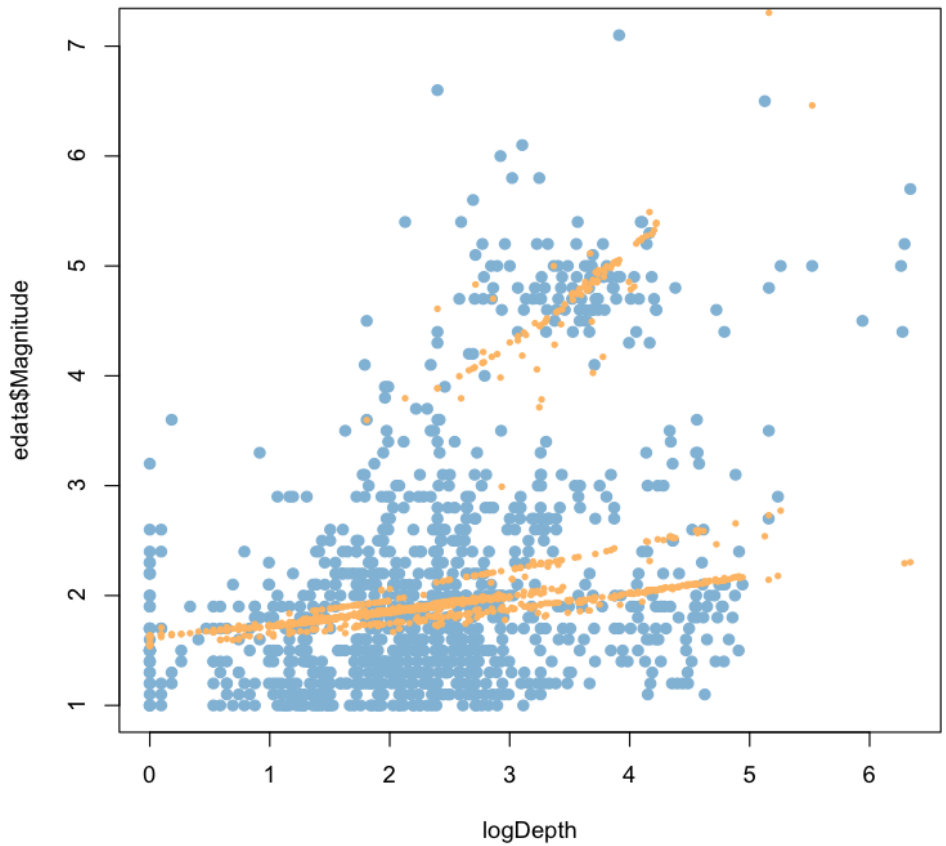
Model Magnitude and log(Depth) + Region

```
>logDepth <- log(edata$Depth + 1)
> glm3 <- glm(edata$Magnitude ~ logDepth + as.factor(edata$Region),family="Gamma")
> plot(logDepth,edata$Magnitude,pch=19,col=cols1[5])
> points(logDepth,glm3$fitted,pch=19,cex=0.5,col=cols1[4])
```



Model Magnitude and log(Depth) + Longitude

```
>logDepth <- log(edata$Depth + 1)
> glm4 <- glm(edata$Magnitude ~ logDepth + edata$Lon,family="Gamma")[2]
> plot(logDepth,edata$Magnitude,pch=19,col=cols1[5])
> points(logDepth,glm4$fitted,pch=19,cex=0.5,col=cols1[6])
```



[2] A compromise would be to treat Longitude as a factor using cut2

Model Magnitude and log(Depth) + Longitude: EEs

The estimating equations for this model are given by:

$$\sum_{i=1}^{n} X_i^T \left(Y_i - \frac{1}{X_i^T \boldsymbol{\beta}} \right) = \mathbf{0}_3$$

where $X_i = [1 \ \text{Log Depth}_i \ \text{Longitude}_i]^T$ and $\boldsymbol{\beta} = [\beta_0 \ \beta_{ld} \ \beta_{long}]^T$
This can be solved with Newton's method (hopefully you did it).

One thing to be careful with for this set of estimating equations is that we need the linear predictor $X_i^T \boldsymbol{\beta}$ to be positive for sensible results.

Sandwich estimation proceeds by the usual formula (which hopefully you worked out for your model!)

Get parameter fits and confidence intervals (non-parametic, sandwich-based)

```
> library(sandwich)
> params = round(cbind(glm4$coeff,glm4$coeff + sqrt(diag(sandwich(glm4))) %o%
qnorm(c(0.025,0.975)))
> rownames(params) = c("Intercept","Log Depth", "Longitude")
> colnames(params) = c("Estimate","Lower","Upper")
> params
Estimate Lower Upper
Intercept 0.4941 0.4619 0.5263
Log Depth -0.0385 -0.0495 -0.0274
Longitude -0.0010 -0.0011 -0.0009
```

First, it is clear that both Log Depth and Longitude are statistically significant at the 0.05 level. Now, how do we interpret each of these parameters?

- The intercept term is interpreted as the inverse of the average magnitude earthquake that occurs on the earth's surface at the same longitude as Greenwich.
- Interpreting the Log Depth term and the Longitude term are somewhat more difficult for the inverse link. The parameters are negative, which suggests that for two earthquakes at the same longitude, on average the trend is that increases in Log Depth are associated with increases in Magnitude (coefficients are in the opposite direction of what you would see for the Poisson model).

Model/Sanity Checking

First lets check that the linear predictor only produces positive values

```
> glm4$coeff[1] + max(logDepth)*glm4$coeff[2] + max(edata$Lon)*glm4$coeff[3]

0.06521122
```

Now see if the coefficient of variation σ/μ is about constant:

```
> dBins = cut2(logDepth,cuts=seq(0,6,by=0.5))
> par(mfcol=c(2,1))
> plot(edata$Magnitude   logDepth,col=cols1[5],pch=19,xlim=c(0,6.8))
> plot(tapply(logDepth,dBins,mean),tapply(glm4$residuals,dBins,sd)/
tapply(glm4$fitted,dBins,mean),

pch=19,xlim=c(0,6.8),col=cols1[1],xlab="Average logDepth",ylab="Average Residual Sd/Average
Fitted Value")
```

Checking Constant CV Assumption