

Regression in the real world

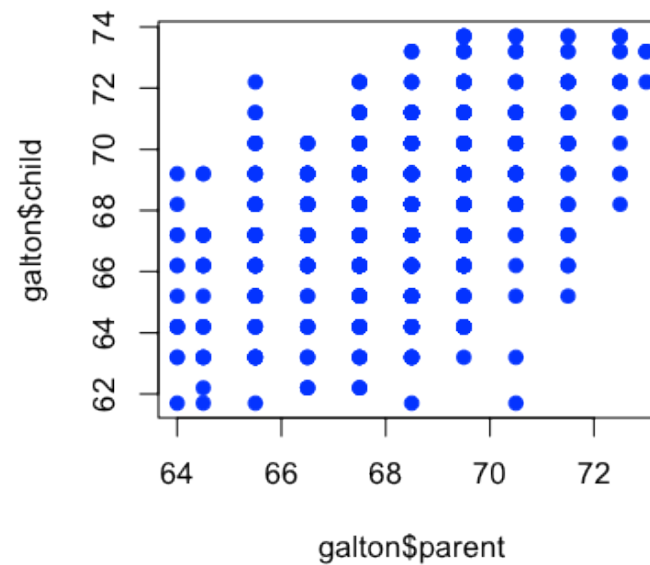
Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Things to pay attention to

- Confounders
- Complicated interactions
- Skewness
- Outliers
- Non-linear patterns
- Variance changes
- Units/scale issues
- Overloading regression
- Correlation and causation

The ideal data for regression

```
library(UsingR); data(galton)
plot(galton$parent,galton$child,col="blue",pch=19)
```



Confounders

Confounder: A variable that is correlated with both the outcome and the covariates

- Confounders can change the regression line
- They can even change the sign of the line
- They can sometimes be detected by careful exploration

Example - Millenium Development Goal 1



GOAL 1 **Eradicate Extreme Poverty and Hunger**

FACT SHEET

TARGETS

1. Halve, between 1990 and 2015, the proportion of people whose income is less than \$1 a day
2. Achieve full and productive employment and decent work for all, including women and young people
3. Halve, between 1990 and 2015, the proportion of people who suffer from hunger

http://www.un.org/millenniumgoals/pdf/MDG_FS_1_EN.pdf

WHO childhood hunger data

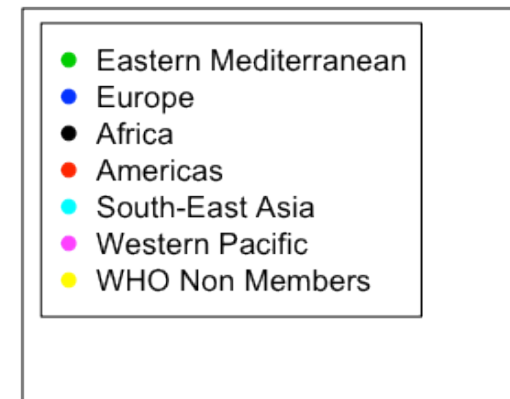
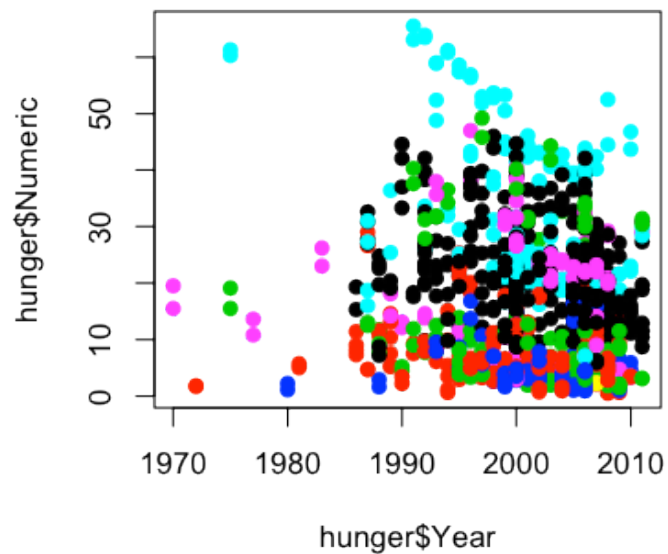
```
download.file("http://apps.who.int/gho/athena/data/GHO/WHOSIS_000008.csv?profile=text&filter=COUNTRY:*;SEX:*","./data/hunger.csv",method="curl")
hunger <- read.csv("./data/hunger.csv")
hunger <- hunger[hunger$Sex!="Both sexes",]
head(hunger)
```

	Indicator	Data.Source	Country	Sex	Year	WHO.region
2	Children aged <5 years underweight (%)	NLIS_312819	Afghanistan	Male	2004	Eastern Mediterranean
4	Children aged <5 years underweight (%)	NLIS_312819	Afghanistan	Female	2004	Eastern Mediterranean
7	Children aged <5 years underweight (%)	NLIS_312361	Albania	Male	2000	Europe
8	Children aged <5 years underweight (%)	NLIS_312361	Albania	Female	2000	Europe
9	Children aged <5 years underweight (%)	NLIS_312879	Albania	Female	2005	Europe
10	Children aged <5 years underweight (%)	NLIS_312879	Albania	Male	2005	Europe
	Display.Value	Numeric	Low	High	Comments	
2	32.7	32.7	NA	NA	NA	
4	33.0	33.0	NA	NA	NA	
7	19.6	19.6	NA	NA	NA	
8	14.2	14.2	NA	NA	NA	
9	5.8	5.8	NA	NA	NA	
10	7.3	7.3	NA	NA	NA	

6/30

Hunger over time by region

```
par(mfrow=c(1,2))  
plot(hunger$Year,hunger$Numeric,col=as.numeric(hunger$WHO.region),pch=19)  
plot(1:10,type="n",xaxt="n",yaxt="n",xlab="",ylab="")  
legend(1,10,col=unique(as.numeric(hunger$WHO.region)),legend=unique(hunger$WHO.region),pch=19)
```



Region correlated with year

```
anova(lm(hunger$Year ~ hunger$WHO.region))
```

Analysis of Variance Table

Response: hunger\$Year

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
hunger\$WHO.region	6	623	103.8	2.25	0.037 *
Residuals	851	39315	46.2		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Region correlated with hunger

```
anova(lm(hunger$Numeric ~ hunger$WHO.region))
```

Analysis of Variance Table

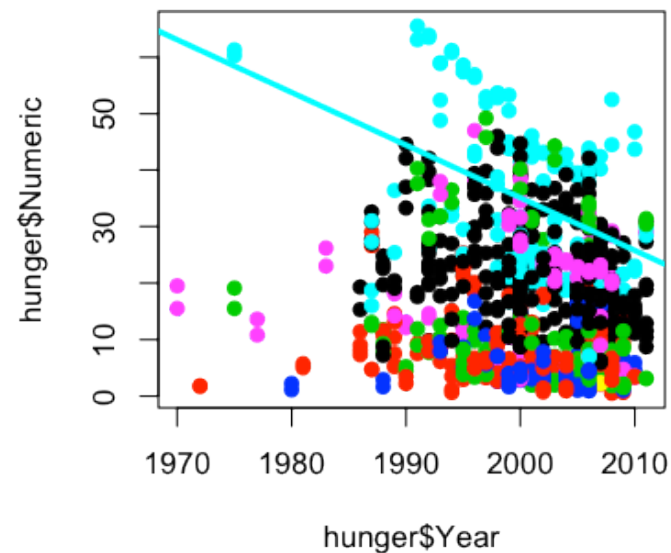
Response: hunger\$Numeric

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
hunger\$WHO.region	6	75042	12507	127	<2e-16 ***
Residuals	851	83853	99		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Including region - a complicated interaction

```
plot(hunger$Year,hunger$Numeric,pch=19,col=as.numeric(hunger$WHO.region))  
lmRegion <- lm(hunger$Numeric ~ hunger$Year + hunger$WHO.region + hunger$Year*hunger$WHO.region )  
abline(c(lmRegion$coeff[1] + lmRegion$coeff[6],lmRegion$coeff[2]+ lmRegion$coeff[12]),col=5,lwd=3)
```



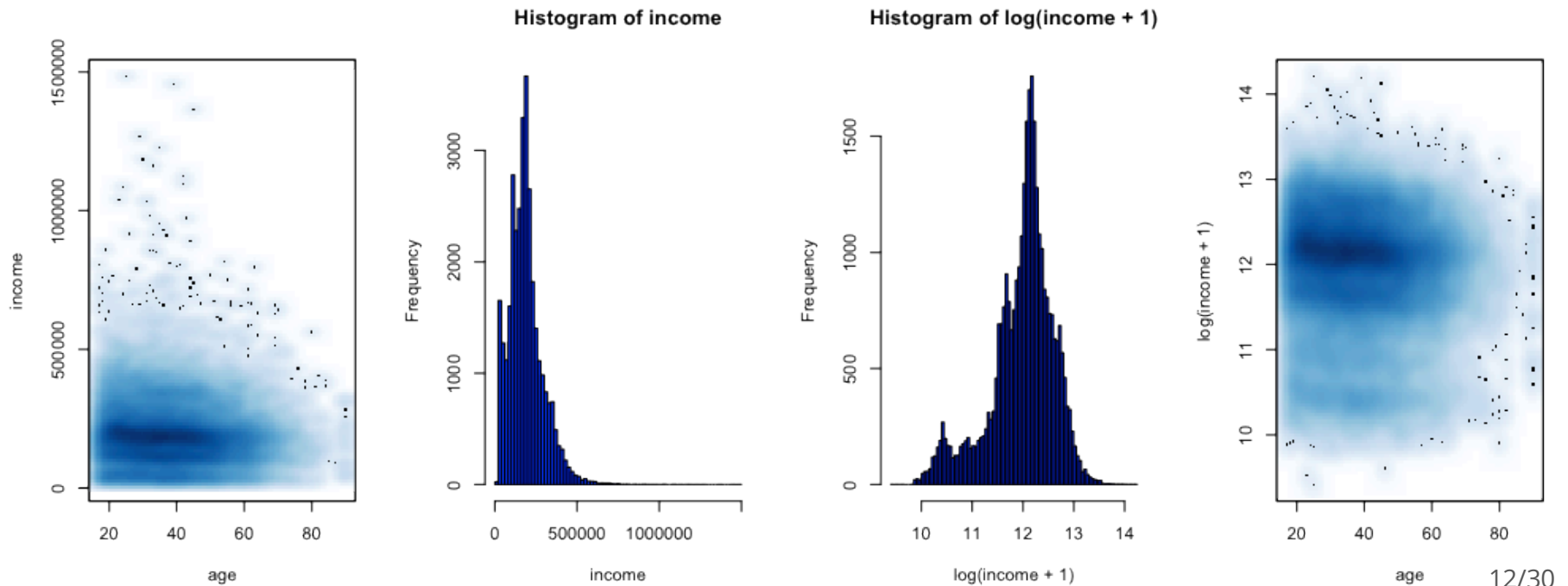
Income data

```
download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data", "./data/i
incomeData <- read.csv("./data/income.csv", header=FALSE)
income <- incomeData[,3]
age <- incomeData[,1]
```

<http://archive.ics.uci.edu/ml/datasets/Census+Income>

Logs to address right-skew

```
par(mfrow=c(1,4))  
smoothScatter(age,income)  
hist(income,col="blue",breaks=100)  
hist(log(income+1),col="blue",breaks=100)  
smoothScatter(age,log(income+1))
```



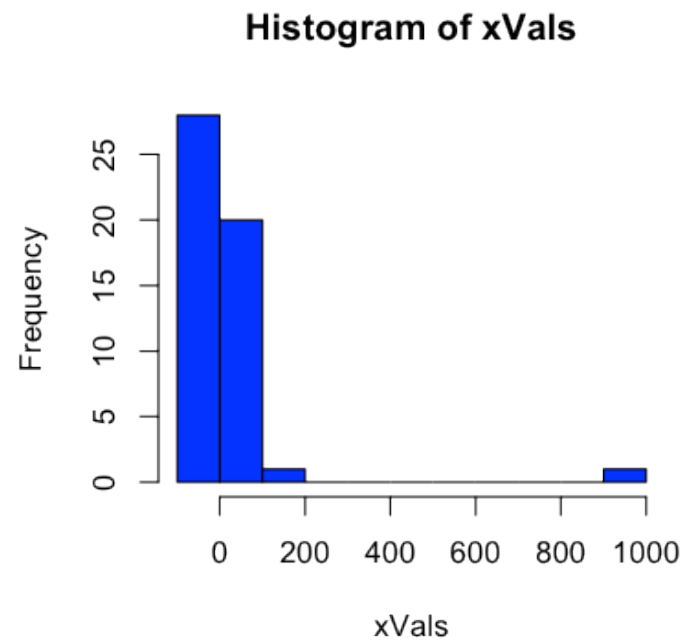
Outliers

"outliers" ... are data points that do not appear to follow the pattern of the other data points.

[A dataset that is 44% outliers](#)

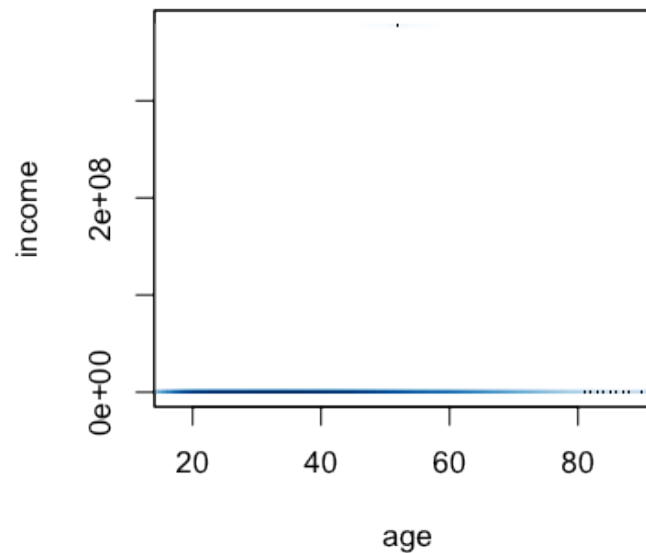
Example - extreme points

```
set.seed(1235)
xVals <- rcauchy(50)
hist(xVals,col="blue")
```



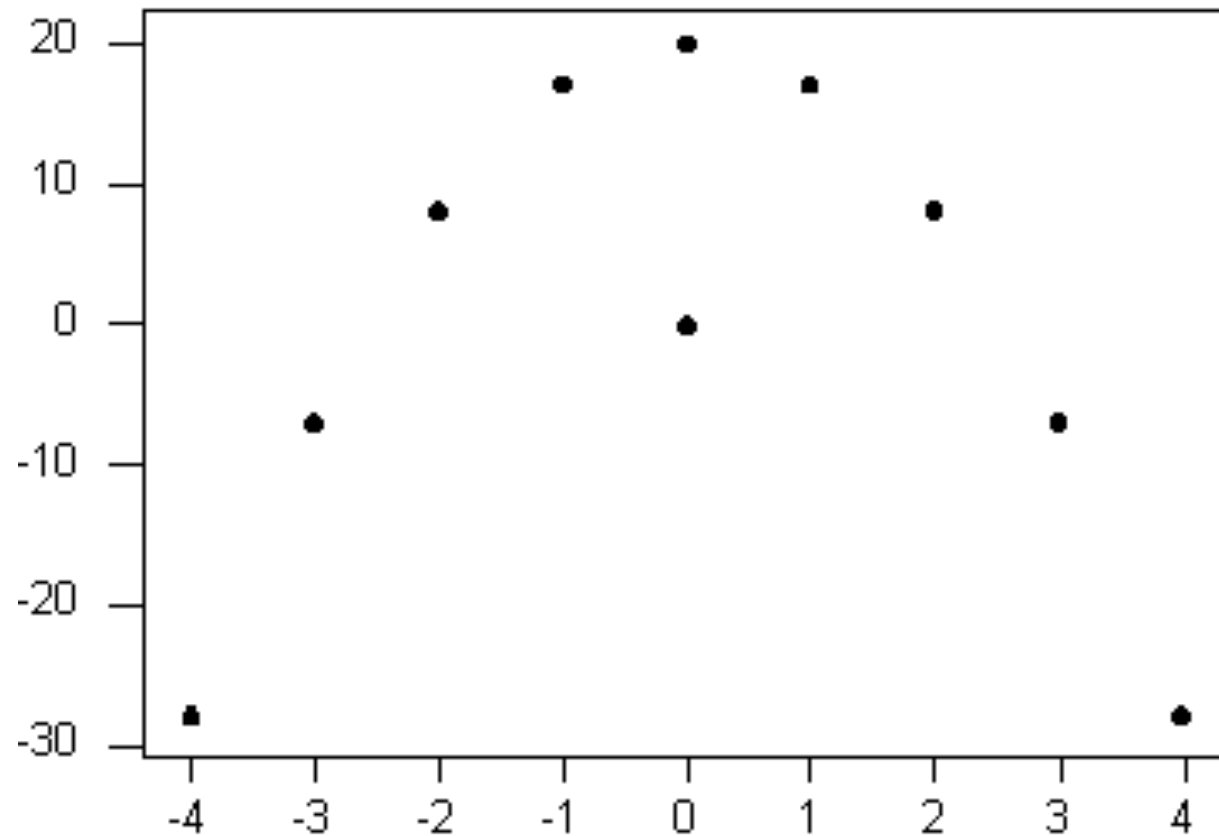
Example - Outliers may be real

```
# Add Tim Cook, CEO of Apple 2011 income  
age <- c(age, 52)  
income <- c(income, 378e6)  
smoothScatter(age, income)
```



<http://www.macworld.com/article/2023491/apple-gives-tim-cook-51-percent-salary-increase.html> ^{15/30}

Example - Does not fit the trend

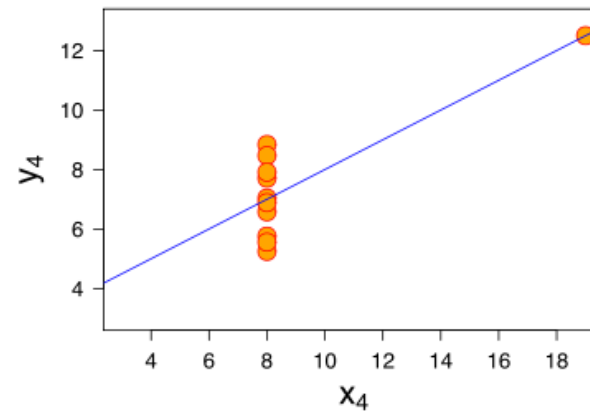
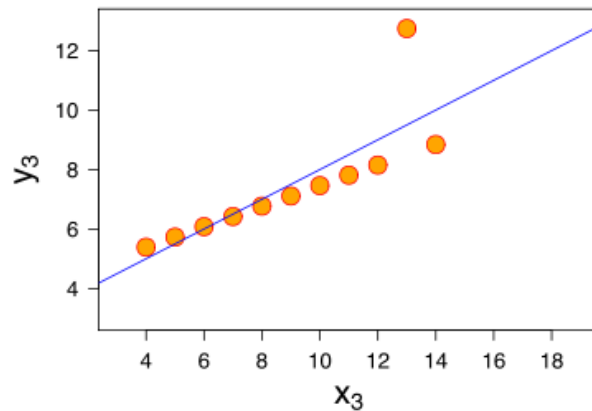
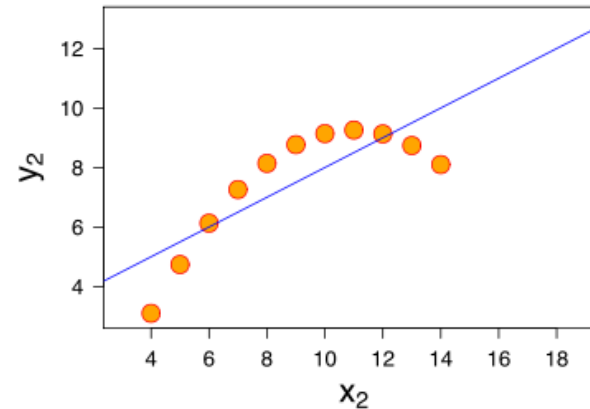
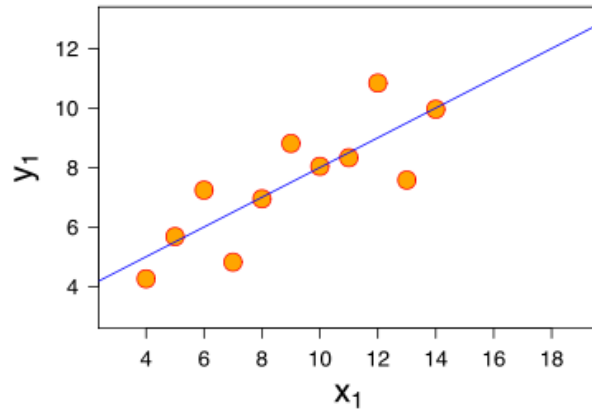


[A dataset that is 44% outliers](#)

Outliers - what you can do

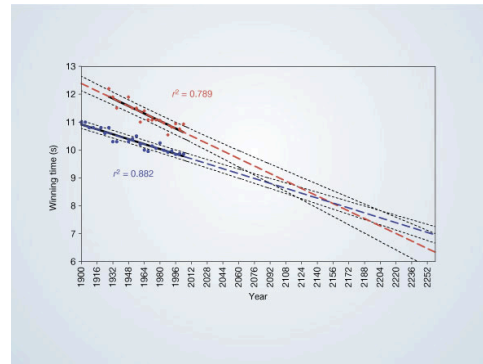
- If you know they aren't real/of interest, remove them (but changes question!)
- Alternatively
 - Sensitivity analysis - is it a big difference if you leave it in/take it out?
 - Logs - if the data are right skewed (lots of outliers)
 - Robust methods - we've been doing averages, but there are more robust approaches ([Robust,rlm](#))

A line isn't always the best summary



http://en.wikipedia.org/wiki/Linear_regression

You can end up saying some pretty silly stuff



http://www.nature.com/nature/journal/v431/n7008/fig_tab/431525a_F1.html

"We are students aged 16–18 in a Texas high school. Our biology teacher Vidya Rajan asked us to comment on the paper by A. J. Tatem and colleagues (Nature 431, 525; 2004); we believe the projection on which it is based is riddled with flaws..."

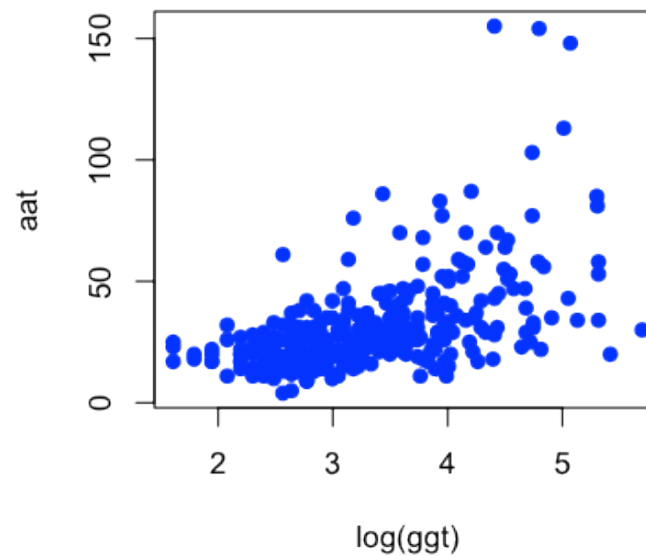
<http://www.nature.com/nature/journal/v432/n7014/full/432147c.html>

"They omit to mention, however, that (according to their analysis) a far more interesting race should occur in about 2636, when times of less than zero seconds will be recorded"

<http://www.nature.com/nature/journal/v432/n7014/full/432147b.html>

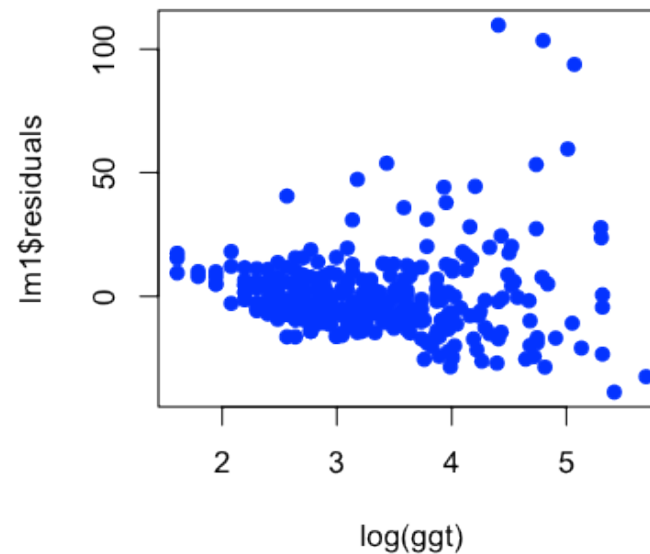
Variance changes

```
bupaData <- read.csv("ftp://ftp.ics.uci.edu/pub/machine-learning-databases/liver-disorders/bupa.dat")  
ggt <- bupaData[,5]; aat <- bupaData[,3]  
plot(log(ggt),aat,col="blue",pch=19)
```



Plot the residuals

```
lm1 <- lm(aat ~ log(ggt))  
plot(log(ggt), lm1$residuals, col="blue", pch=19)
```

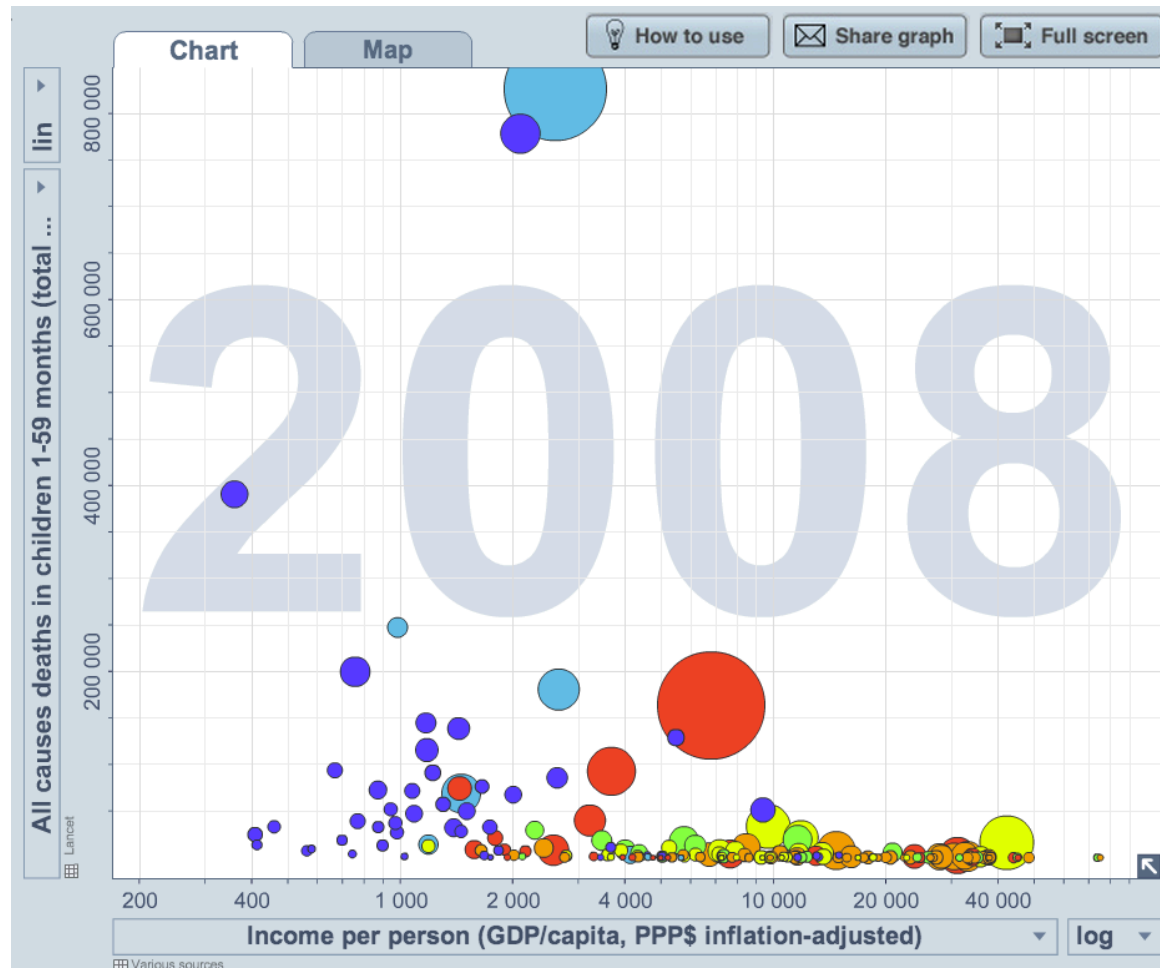


[Power \(a.k.a. Box-Cox\) transform](#)

Changing variance - what you can do

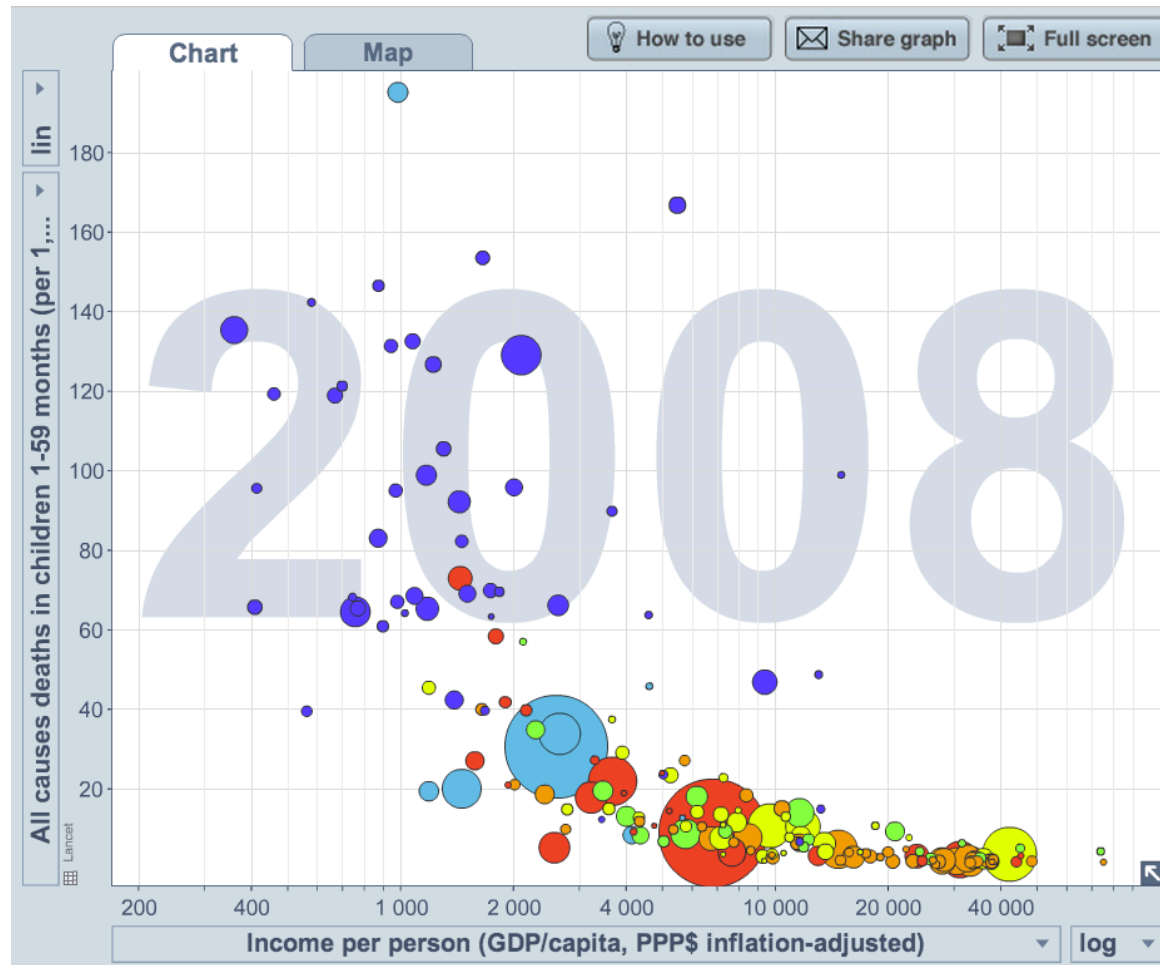
- There is a long literature on this problem (heteroskedasticity)
- A few examples
 - [Box-Cox Transform](#)
 - [Variance stabilizing transform](#)
 - [Weighted least squares](#)
 - [Huber-white standard errors](#)

Variation in units



All Deaths

Relative units



Per 1000 Deaths

24/30

When there is variation in units

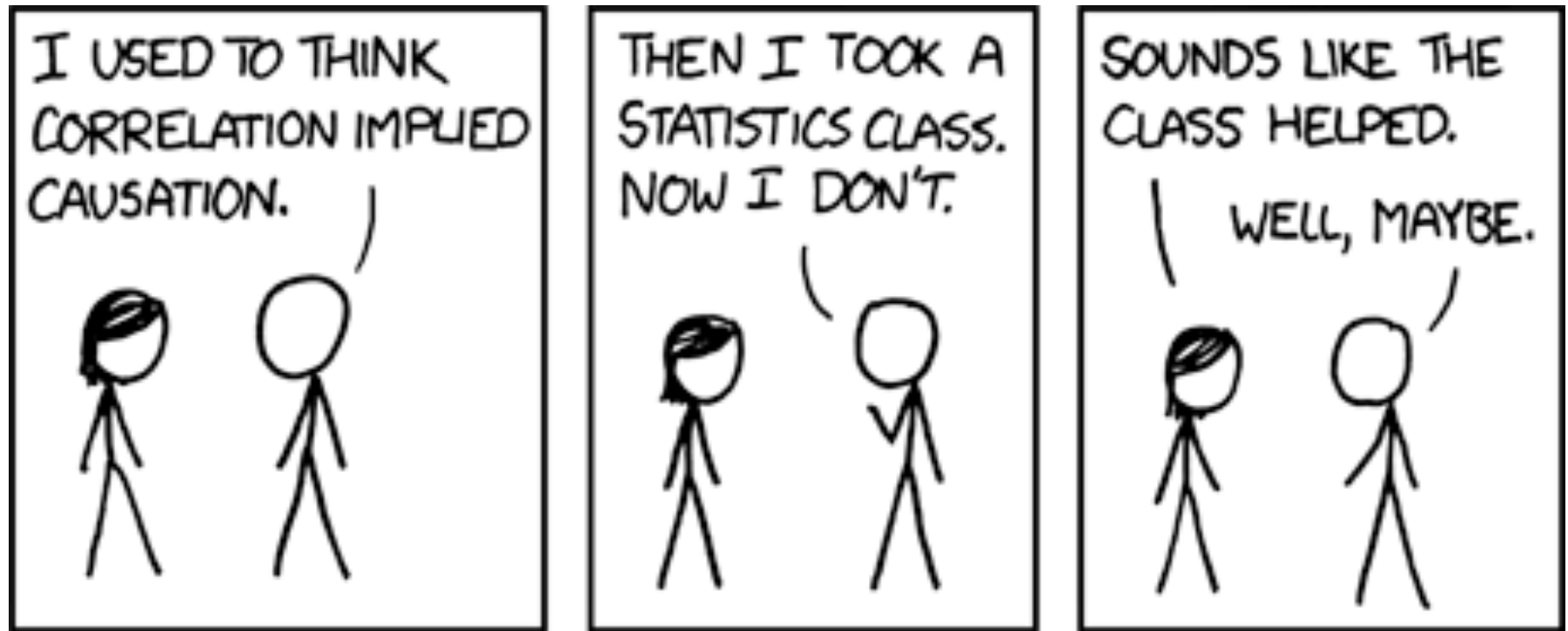
- Standardize, but keep track
 - Affects model fits
 - Affects interpretation
 - Affects inference

Overloading regression

$$AMR_{it} = \alpha + \beta_1 PRIV_{it} + \beta_2 GDP_{it} + \beta_3 LIB_{it} + \beta_4 TRADE_{it} + \beta_5 DEM_{it} + \beta_6 WAR_{it} + \beta_7 DEP_{it} + \beta_8 URBAN_{it} + \beta_9 EDUC_{it} + \mu_i + \epsilon_{it}$$

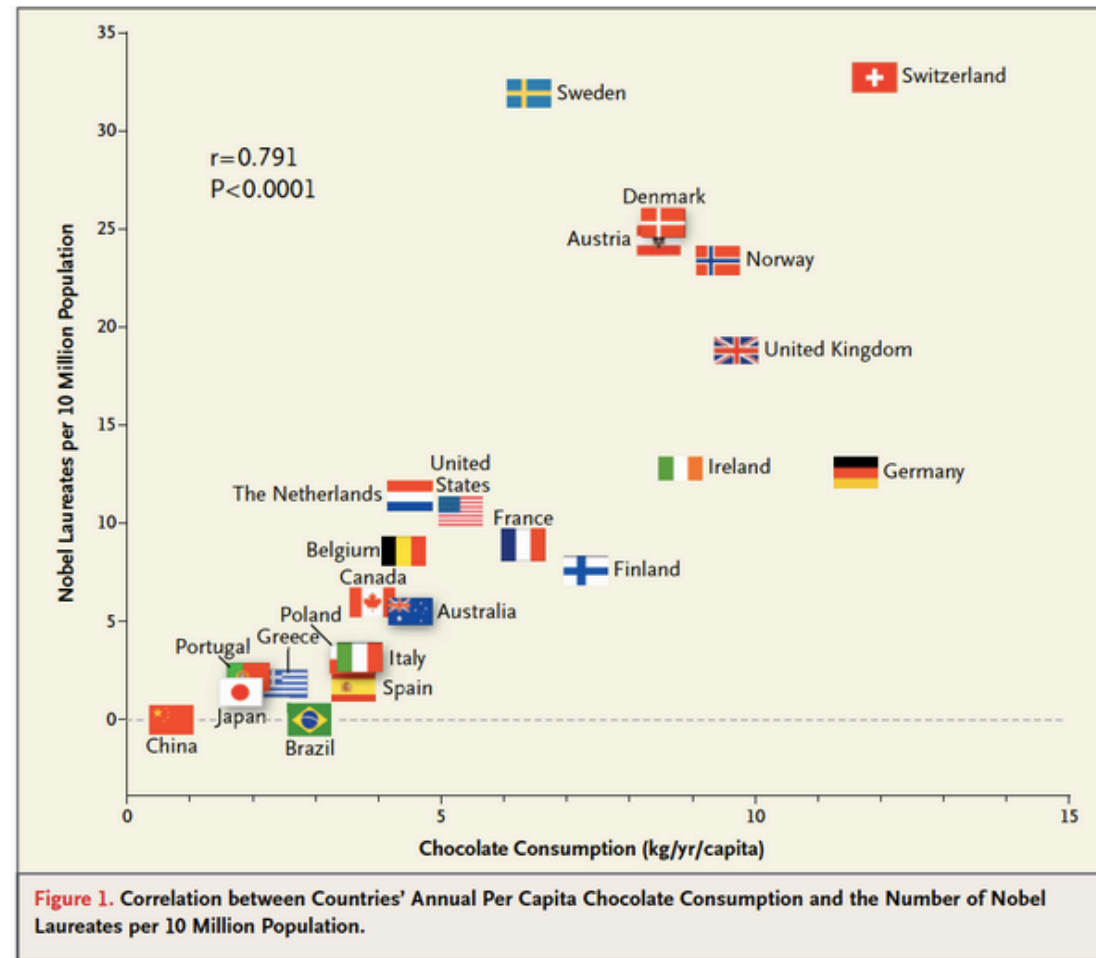
<http://bit.ly/YiB5Um> <http://wmbriggs.com/blog/?p=7026>

Correlation and Causation



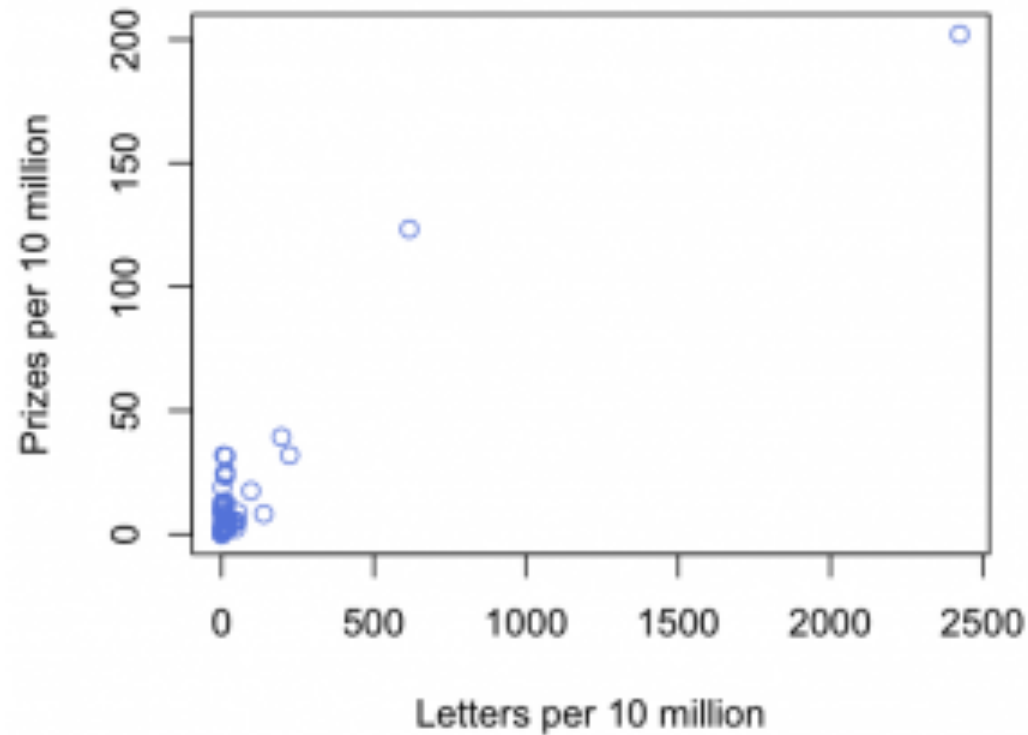
<http://xkcd.com/552/>

Even when looking for associations



<http://www.nejm.org/doi/full/10.1056/NEJMon1211064>

Again, it can get silly



Correlation vs. Causation

- Use caution when interpreting regression results
- Be critical of surprising associations
- Consider alternative explanations