Multiple regression

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

Key ideas

- Regression with multiple covariates
- Still using least squares/central limit theorem
- Interpretation depends on all variables

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

http://apps.who.int/gho/athena/data/GHO/WHOSIS_000008.csv?profile=text&filter=COUNTRY:;*SEX:*

3/16

WHO childhood hunger data

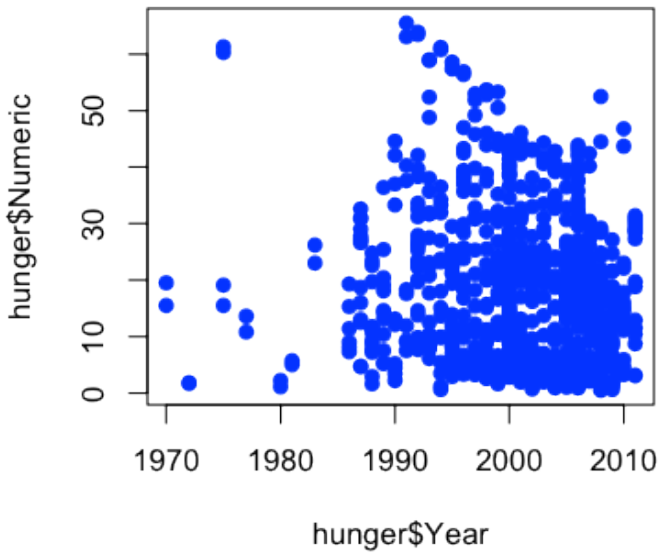
```
download.file("http://apps.who.int/gho/athena/data/GHO/WHOSIS_000008.csv?profile=text&filter=COUNTR
hunger <- read.csv("./data/hunger.csv")
hunger <- hunger[hunger$Sex!="Both sexes",]
head(hunger)
```

```
                                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
```

4/16

Plot percent hungry versus time

```
lm1 <- lm(hunger$Numeric ~ hunger$Year)
plot(hunger$Year,hunger$Numeric,pch=19,col="blue")
```



5/16

Remember the linear model

$$Hu_i = b_0 + b_1 Y_i + e_i$$

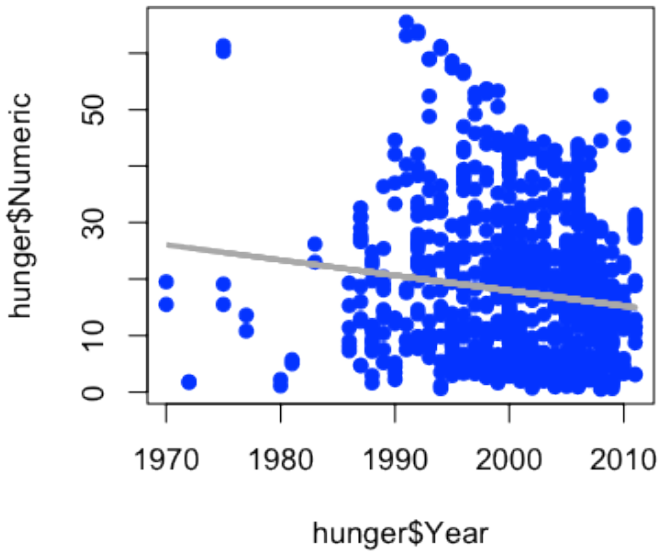
b_0 = percent hungry at Year 0

b_1 = decrease in percent hungry per year

e_i = everything we didn't measure

Add the linear model

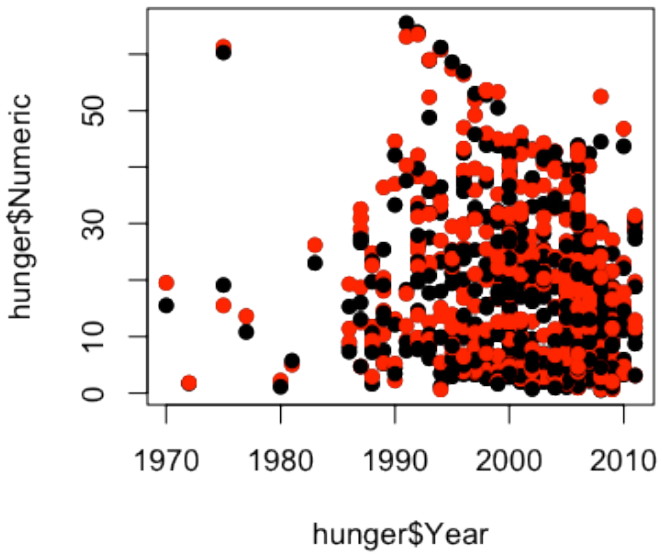
```
lm1 <- lm(hunger$Numeric ~ hunger$Year)
plot(hunger$Year,hunger$Numeric,pch=19,col="blue")
lines(hunger$Year,lm1$fitted,lwd=3,col="darkgrey")
```



7/16

Color by male/female

```
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=((hunger$Sex=="Male")*1+1))
```



8/16

Now two lines

$$HuF_i = bf_0 + bf_1 YF_i + ef_i$$

bf_0 = percent of girls hungry at Year 0

bf_1 = decrease in percent of girls hungry per year

ef_i = everything we didn't measure

$$HuM_i = bm_0 + bm_1 YM_i + em_i$$

bm_0 = percent of boys hungry at Year 0

bm_1 = decrease in percent of boys hungry per year

em_i = everything we didn't measure

Color by male/female

```
lmM <- lm(hunger$Numeric[hunger$Sex=="Male"] ~ hunger$Year[hunger$Sex=="Male"])
lmF <- lm(hunger$Numeric[hunger$Sex=="Female"] ~ hunger$Year[hunger$Sex=="Female"])
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=((hunger$Sex=="Male")*1+1))
lines(hunger$Year[hunger$Sex=="Male"],lmM$fitted,col="black",lwd=3)
lines(hunger$Year[hunger$Sex=="Female"],lmF$fitted,col="red",lwd=3)
```

Two lines, same slope

$$Hu_i = b_0 + b_1 \mathbb{1}(Sex_i = \text{" } Male \text{ "}) + b_2 Y_i + e_i^*$$

b_0 - percent hungry at year zero for females

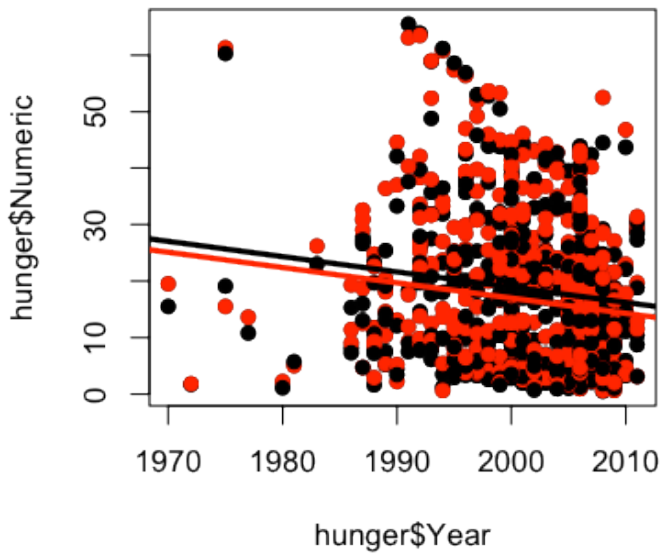
$b_0 + b_1$ - percent hungry at year zero for males

b_2 - change in percent hungry (for either males or females) in one year

e_i^* - everything we didn't measure

Two lines, same slope in R

```
lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex)
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=((hunger$Sex=="Male")*1+1))
abline(c(lmBoth$coeff[1],lmBoth$coeff[2]),col="red",lwd=3)
abline(c(lmBoth$coeff[1] + lmBoth$coeff[3],lmBoth$coeff[2] ),col="black",lwd=3)
```



12/16

Two lines, different slopes (interactions)

$$Hu_i = b_0 + b_1 \mathbb{1}(Sex_i = \text{" } Male \text{ "}) + b_2 Y_i + b_3 \mathbb{1}(Sex_i = \text{" } Male \text{ "}) \times Y_i + e_i^+$$

b_0 - percent hungry at year zero for females

$b_0 + b_1$ - percent hungry at year zero for males

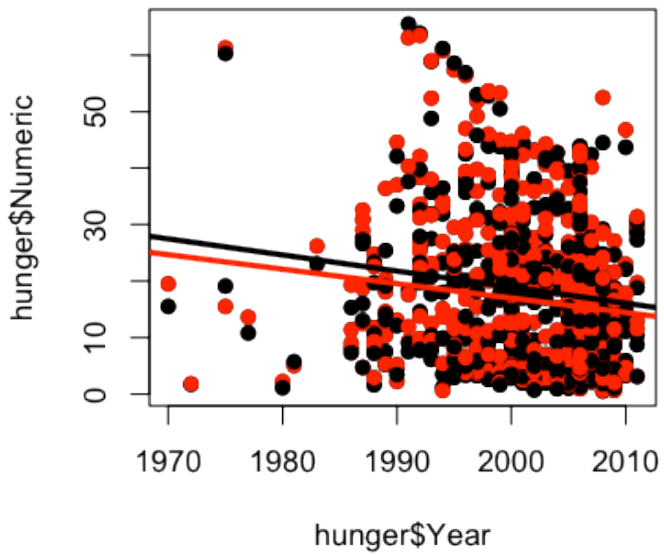
b_2 - change in percent hungry (females) in one year

$b_2 + b_3$ - change in percent hungry (males) in one year

e_i^+ - everything we didn't measure

Two lines, different slopes in R

```
lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex + hunger$Sex*hunger$Year)
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=((hunger$Sex=="Male")*1+1))
abline(c(lmBoth$coeff[1],lmBoth$coeff[2]),col="red",lwd=3)
abline(c(lmBoth$coeff[1] + lmBoth$coeff[3],lmBoth$coeff[2] +lmBoth$coeff[4]),col="black",lwd=3)
```



14/16

Two lines, different slopes in R

```
summary(lmBoth)
```

```
Call:
lm(formula = hunger$Numeric ~ hunger$Year + hunger$Sex + hunger$Sex *
    hunger$Year)

Residuals:
   Min     1Q Median     3Q    Max
-25.11 -11.55  -2.12   7.02  46.22

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               529.4033   190.8185    2.77   0.0057 **
hunger$Year                -0.2562     0.0954   -2.69   0.0074 **
hunger$SexMale             59.5912   269.8581    0.22   0.8253
hunger$Year:hunger$SexMale -0.0288     0.1349   -0.21   0.8309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions for continuous variables

$$Hu_i = b_0 + b_1 In_i + b_2 Y_i + b_3 In_i \times Y_i + e_i^+$$

b_0 - percent hungry at year zero for children with whose parents have no income

b_1 - change in percent hungry for each dollar of income in year zero

b_2 - change in percent hungry in one year for children whose parents have no income

b_3 - increased change in percent hungry by year for each dollar of income - e.g. if income is $10,000, then change in percent hungry in one year will be

$$b_2 + 1e4 \times b_3$$

e_i^+ - everything we didn't measure

Lot's of care/caution needed!

16/16