Today

- bivariate correlation
- bivariate regression
- multiple regression

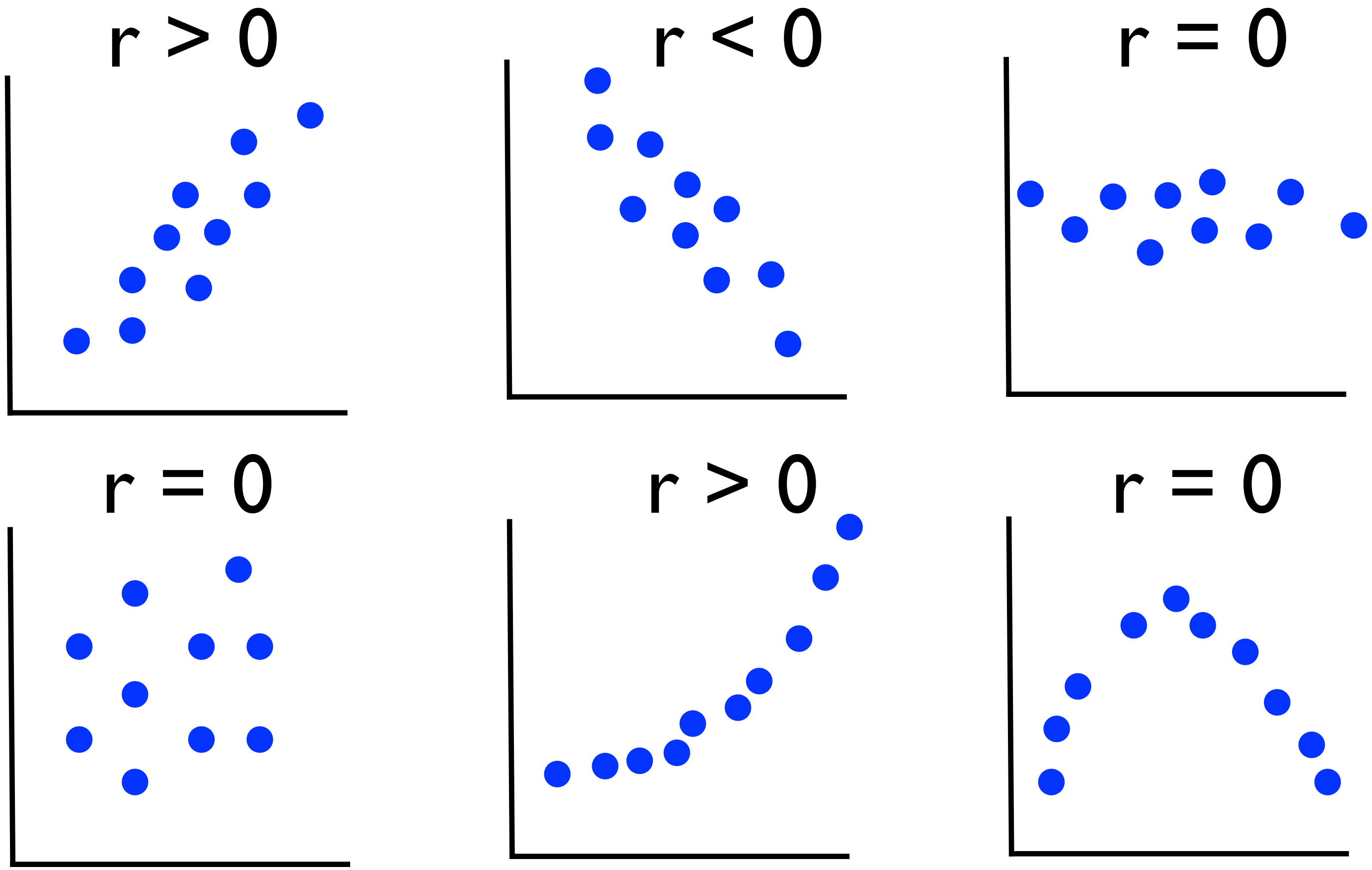
Bivariate Correlation

- Pearson product-moment correlation (r)
- assesses nature and strength of the **linear** relationship between two continuous variables

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

- r^2 represents proportion of variance shared by the two variables
- e.g. r=0.663, r^2=0.439: X and Y share 43.9% of the variance in common

Bivariate Correlation



remember: *r* measures **linear** correlation

Significance Tests

- we can perform significance tests on r
- H0: (population) r = 0;
  H1: (population) r not equal to 0 (two-tailed)
  H1: (population) $r < 0$ (or >0) : one-tailed
- sampling distribution of r
- IF we were to randomly draw two samples from two populations that were not correlated at all, what proportion of the time would we get a value of r as as extreme as we observe?
- if $p < .05$ we reject H0

Significance Tests

- We can perform an F-test:
  df = (1,N-2)

$$F = \frac{r^2(N-2)}{1-r^2}$$

- or we could also do a t-test:
  df = N-2

$$t = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$$

- so for example, if we have an observed r = 0.663 based on a sample of 10 (X,Y) pairs
  - Fobs = 6.261
  - Fcrit(1,8,0.05) = 5.32 (or compute p = 0.018)
  - therefore reject H0

Significance Tests

- be careful! statistical significance does not equal scientific significance
- e.g. let's say we have 112 data points
  we compute $r = 0.2134$
  we do an F-test: $F_{obs}(1,110) = 5.34$, $p < .05$
  reject H0! we have a "significant" correlation
- if $r=0.2134$, $r^2 = 0.046$
  only 4.6% of the variance is shared
  between X and Y
  95.4% of the variance is NOT shared
- *H0 is that $r = 0$, not that r is large (not that r is* ***significant****)*

Bivariate Regression

- X, Y continuous variables
- Y is considered to be dependent on X
- we want to predict a value of Y, given a value of X
- e.g. Y is a person's weight, X is a person's height

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

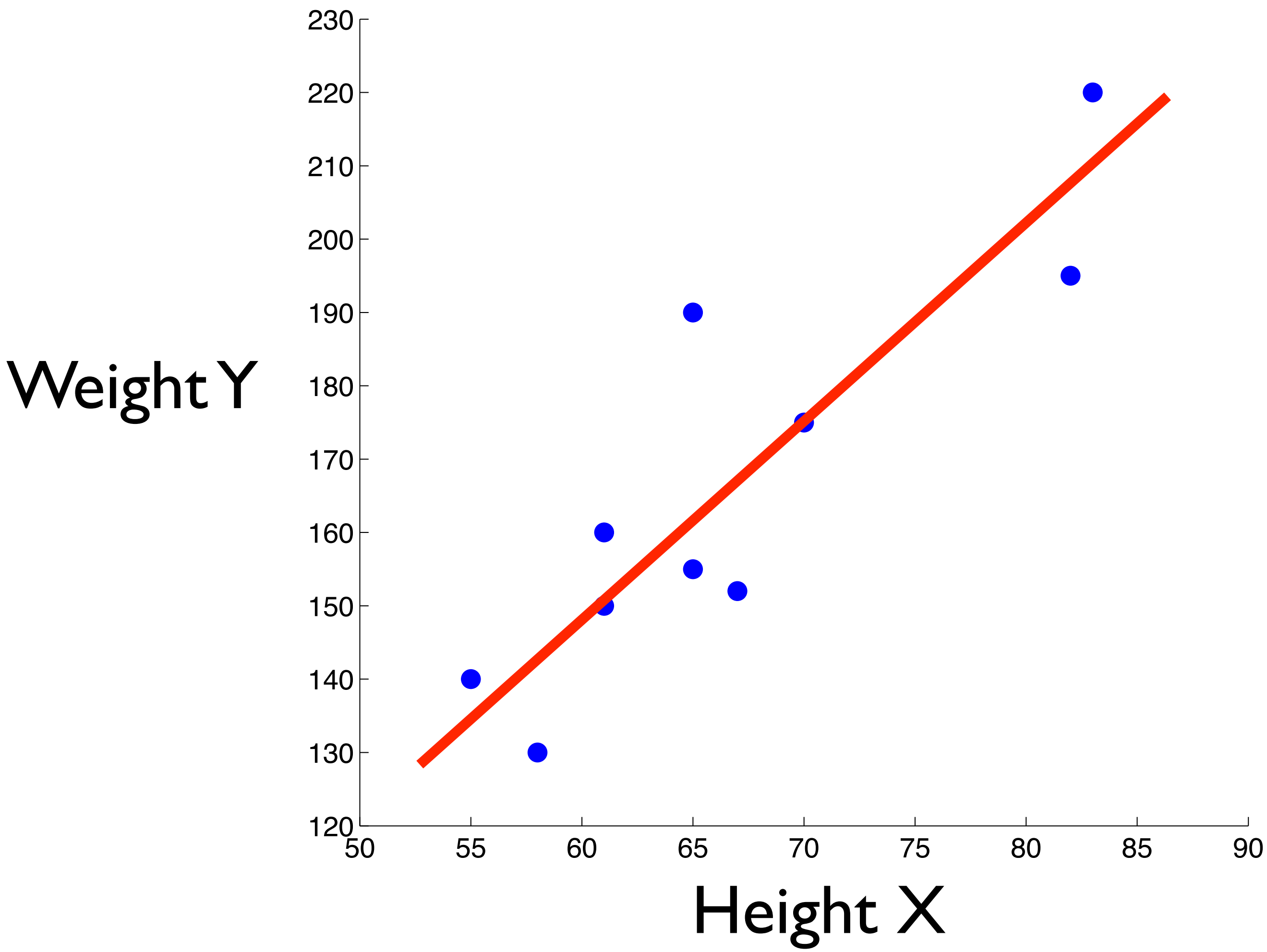
- estimate of Y, Yhat_i, is equal to a constant (beta_0) plus another constant (beta_1) times the value of X
- this is the equation for a straight line
- beta_0 is the Y-intercept, beta_1 is the slope

Bivariate Regression

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

- we want to predict Y given X
- we are modelling Y using a linear equation

Height (X)	Weight (Y)
55	140
61	150
67	152
83	220
65	190
82	195
70	175
58	130
65	155
61	160



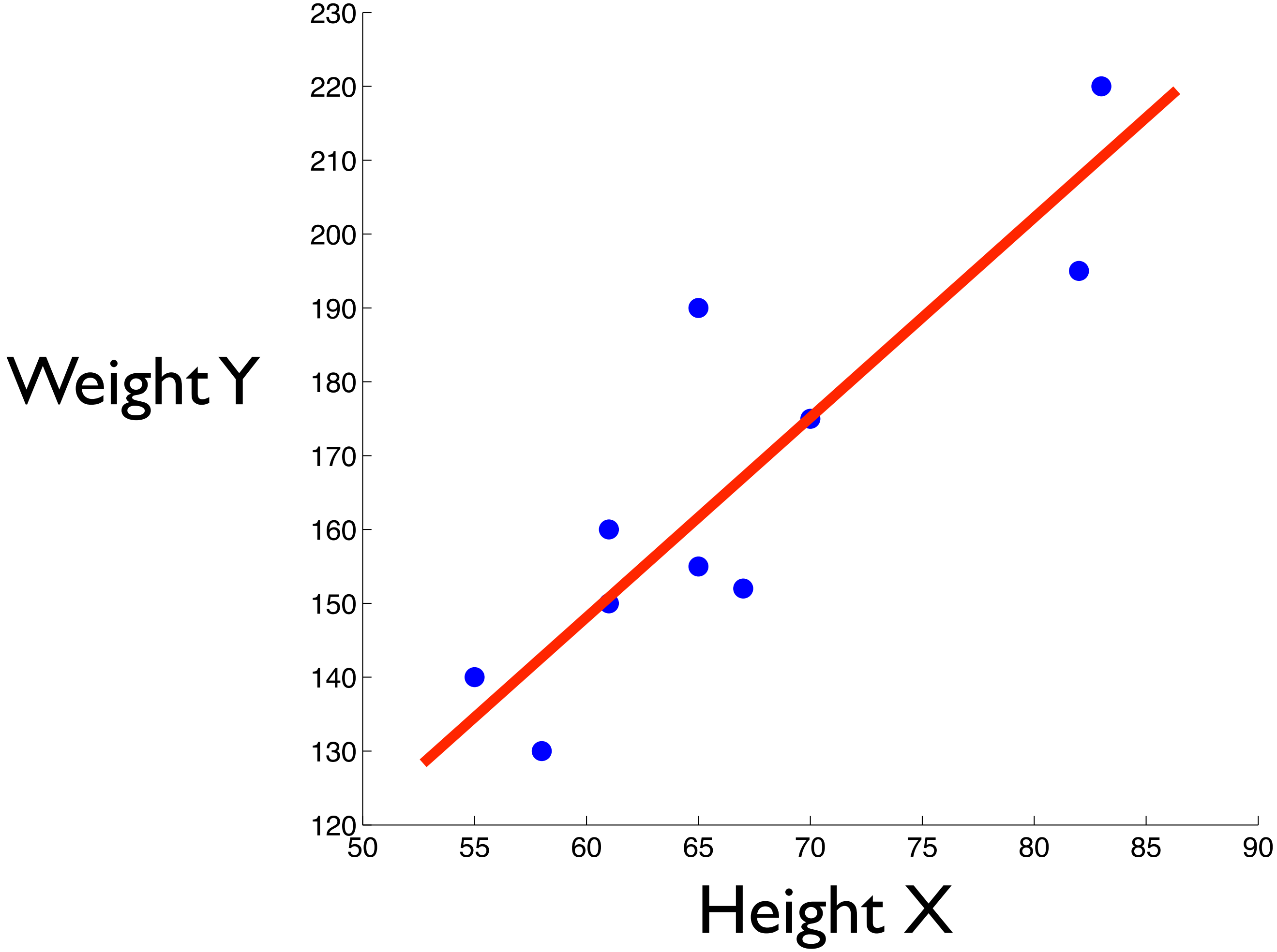
$$\beta_0 = -7.2$$
$$\beta_1 = 2.6$$

Bivariate Regression

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

Height (X)	Weight (Y)
55	140
61	150
67	152
83	220
65	190
82	195
70	175
58	130
65	155
61	160

- slope means that every inch in height is associated with 2.6 pounds of weight



$$\beta_0 = -7.2$$

$$\boxed{\beta_1 = 2.6}$$

Bivariate Regression

- How do we estimate the coefficients beta_0 and beta_1?
- for bivariate regression there are formulas:

$$\beta_1 = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

- These formulas estimate beta_0 and beta_1 according to a least-squares criterion
- they are **the** two beta values that minimize the sum of squared deviations between the estimated values of Y (the line of best fit) and the actual values of Y (the data)

Bivariate Regression

- How good is our line of best fit?
- common measure is "Standard Error of Estimate"

$$SE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N - 2}}$$

- N is number of (X,Y) pairs of data
- SE gives a measure of the typical prediction error in units of Y
- e.g. in our height/weight data
  - SE = sqrt(1596 / 8) = 14.1 lbs

Bivariate Regression

- we can use SE to generate confidence intervals for our estimated values

$$\hat{Y} = (\beta_0 + \beta_1 X) \pm 1.96SE$$

- so for example if height = 72 inches, predicted weight is
  - -7.2 + 2.6*72 = 180 pounds, +/- 1.96(14.1)
  - = 180 +/- 27.6 pounds
- this means that if we take repeated samples from the population, and recompute the regression line, that 95% of the time we will find a confidence interval that will contain the true population mean weight of a 72 inch tall individual, within the endpoints of the CI of that sample
- obviously SE and thus CI depends on size of sample (N)

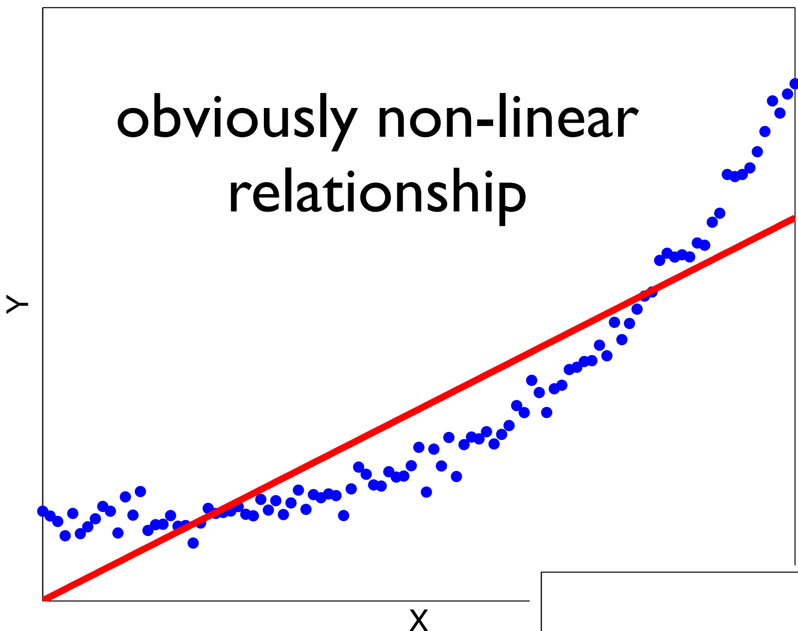
Bivariate Regression

- another measure of fit: r^2
- r^2 gives the proportion of variance accounted for
- e.g. r^2 = 0.58 means that 58% of the variance in Y is accounted for by X
- r^2 is bounded by [0,1]

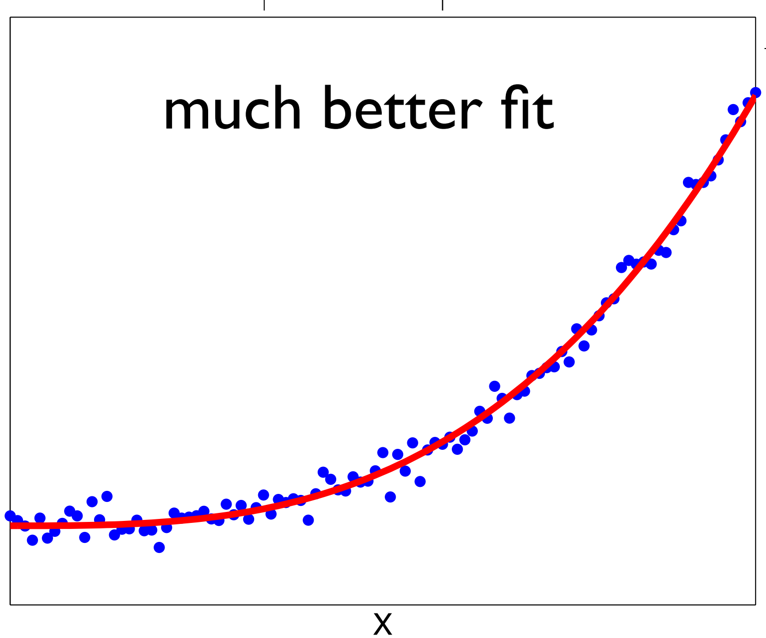
$$r^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

Linear Regression with Non-Linear Terms

$$Y = \beta_0 + \beta_1 X$$

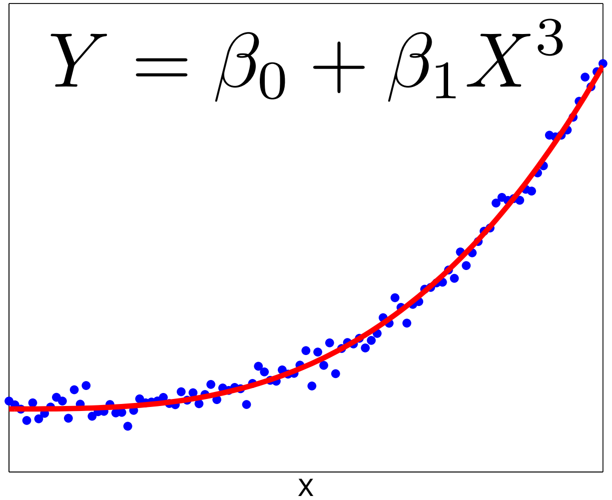


$$Y = \beta_0 + \beta_1 X^2$$



$$Y = \beta_0 + \beta_1 X^3$$

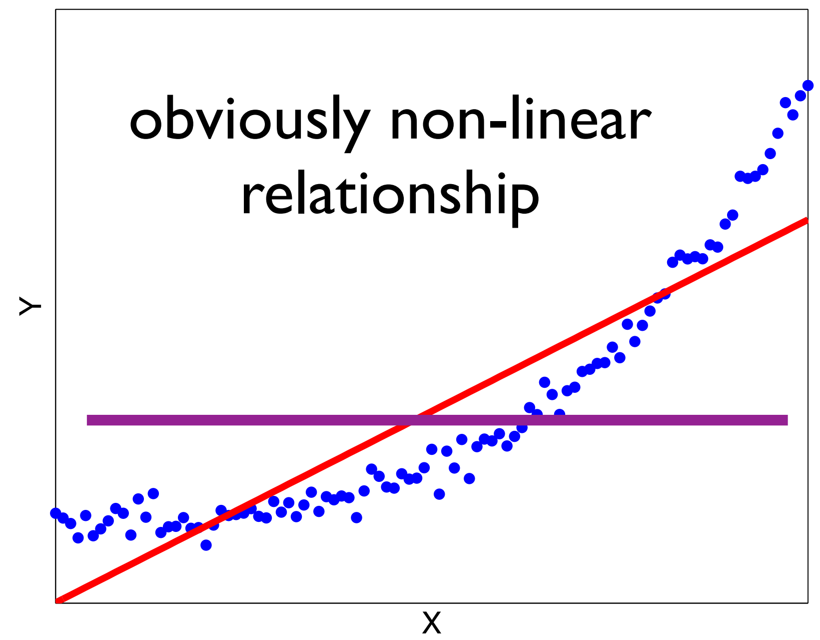
Linear Regression with Non-Linear Terms



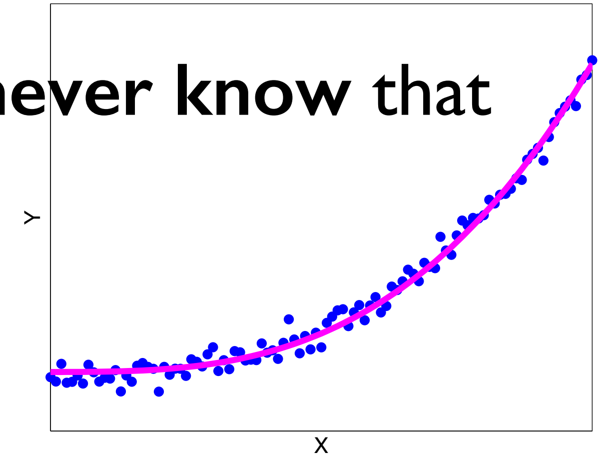
- How do we do this?
- Just create a new variable X^3
- then perform linear regression using that instead of X
- you will get your beta coefficients and r^2
- you can generate predicted values of Y if you want

Always plot your data

$$Y = \beta_0 + \beta_1 X$$



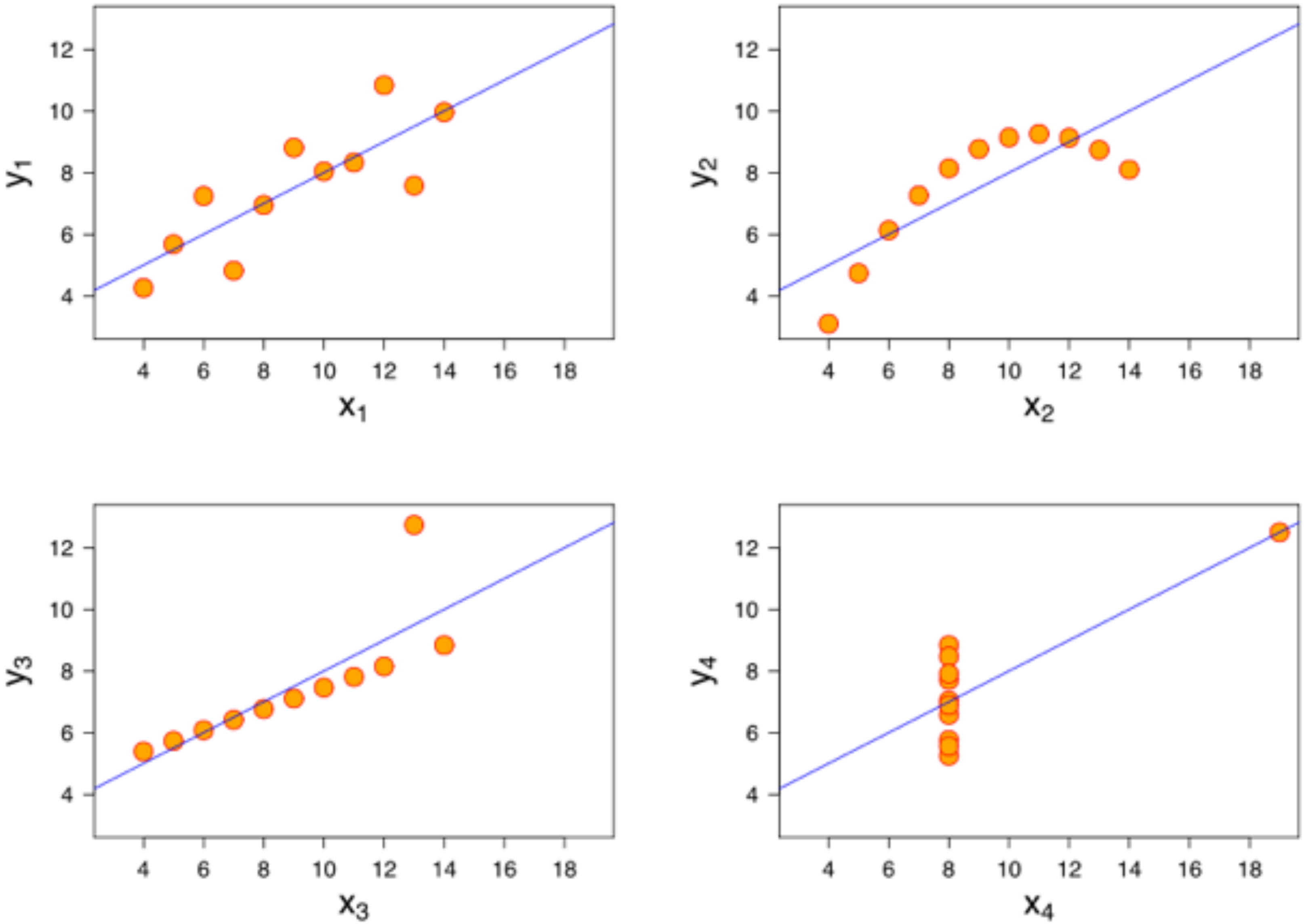
- this poor fitting regression line gives the following F-test:
- $F(1,99)=266.2, p < .001$
- $r^2 = 0.85$
- so we have accounted for 85% of the variance in Y using a straight line
- is this good enough? what is H0? ($y = B0$)
- if you never plotted the data **you would never know** that you can do a **LOT** better
- with $Y = B0 + B1(X^3)$ we get $r^2 = 0.99$



Anscombe's quartet

- four datasets that have nearly identical simple statistical properties, yet appear very different when graphed
- each dataset consists of eleven (x,y) points
- constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties
- http://en.wikipedia.org/wiki/Anscombe's_quartet

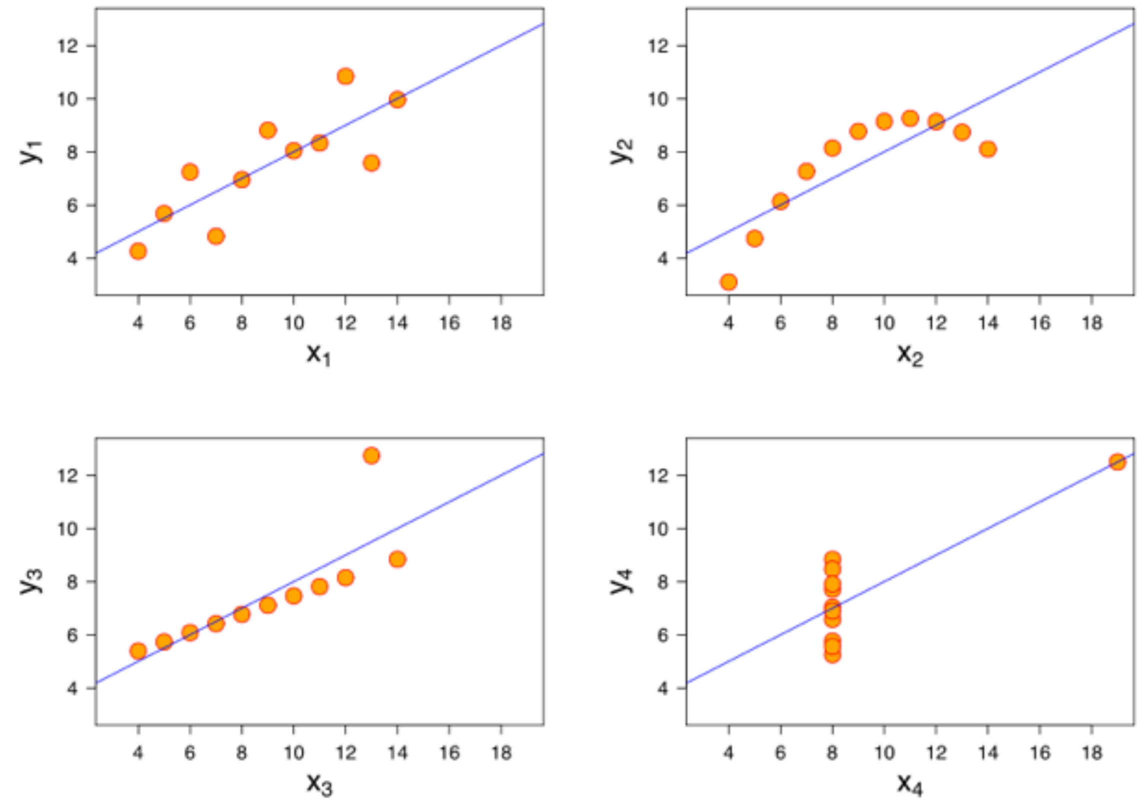
Anscombe's quartet



Anscombe's quartet

in all 4 cases:

- mean(x) = 9
- var(x) = 11
- mean(y) = 7.50
- var(y) = 4.122 or 4.127
- cor(x,y) = 0.816
- regression:
  y = 3.00 + 0.500 (x)



Multiple Regression

- same idea as bivariate regression
- we want to predict values of a continuous variable Y
- but instead of basing our prediction on a single variable X,
- we will use several independent variables X1 .. Xk
- the linear model is:

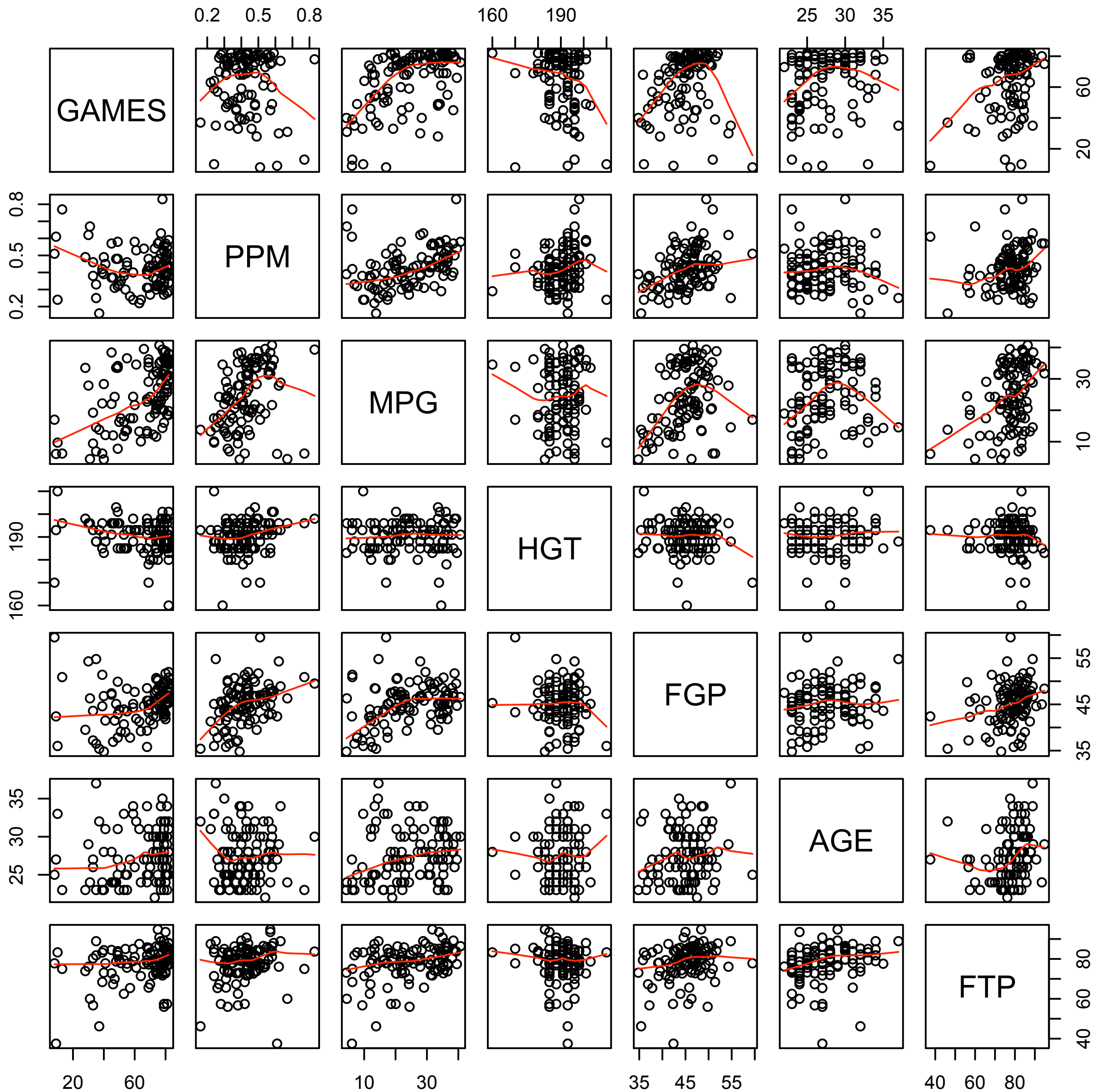
$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

- betas are constants, X1, ..., Xk are predictor variables
- beta weights are found which minimize the total sum of squared error between the predicted and actual Y values

An Example

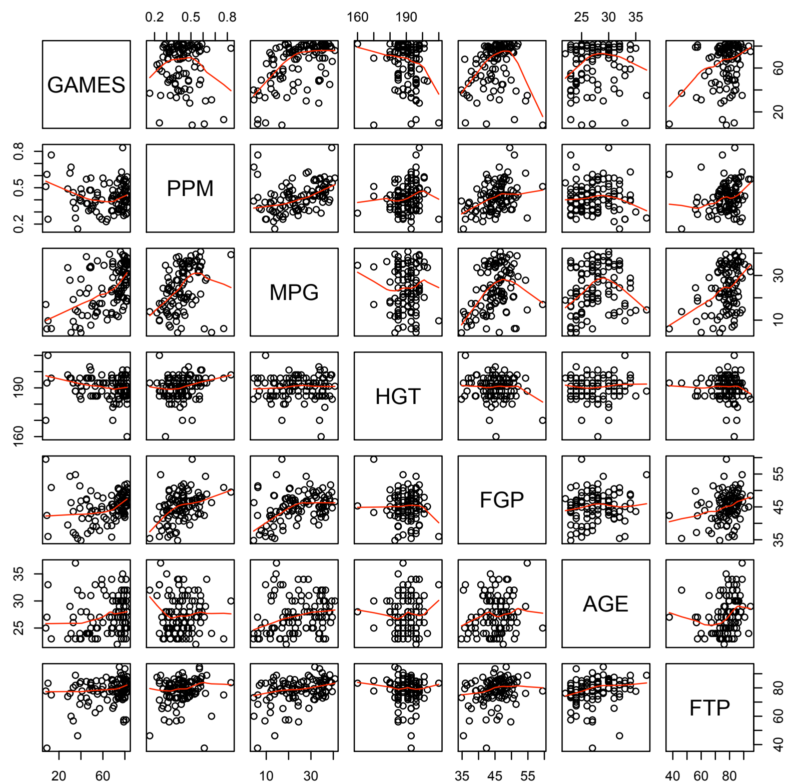
- basketball data
  http://www.gribblelab.org/stats/data/bball.csv
- data from 105 NBA players
  - # games played last season
  - points scored per minute
  - minutes played per game
  - height
  - field goal percentage
  - age
  - free throw percentage
- You are the new coach. You want to develop a model that will let you predict points scored per minute based on the other 6 variables

```
> mydata <- read.table("http://www.gribblelab.org/stats/data/bball.csv", header=T, sep=",")
> plot(mydata)
```



Questions answered by Multiple Regression

- What is the best single predictor?
- What is the best equation (model)?
- Does a certain variable add significantly to the predictive power?



What is the best single predictor?

- simply obtain the bivariate correlations between the dependent variable (Y) and each of the individual predictor variables (X1-X6)
- which predictors have a significant correlation?
- predictor with the maximum (absolute) correlation coefficient is the best single predictor
- (note largest r can be negative)

points per minute PPM vs:		
predictor	r	p
age	-0.0442	0.654
field goal %	0.4063	0.00...
free throw %	0.1655	0.092
games/season	-0.0598	0.544
height	0.2134	0.029
minutes/game	0.3562	0.00...

What is the best model?

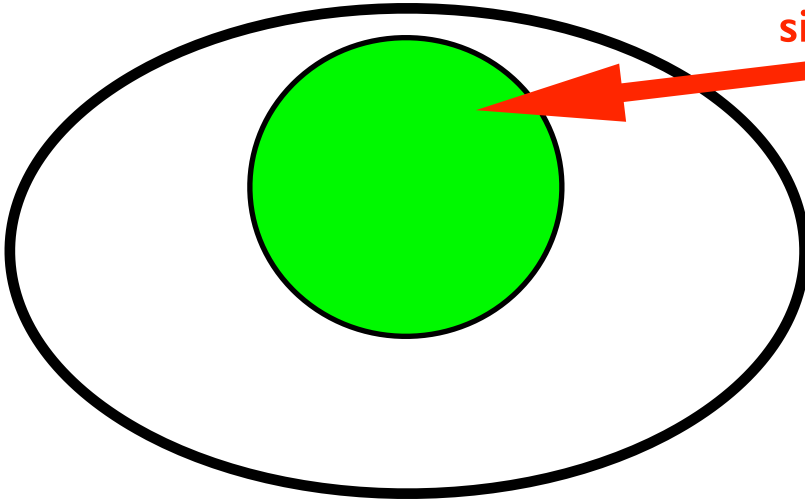
- 3 ways to do this:

- forward regression
- backward regression
- stepwise regression

Forward Regression

1. start with no IVs in the equation
2. check to see if any IVs significantly predict DV
3. if no, stop
   if yes, add the best IV and go to step 4
4. check to see if any remaining IVs predict a significant unique amount of variance
5. if no, stop
   if yes, add the best and go to step 4

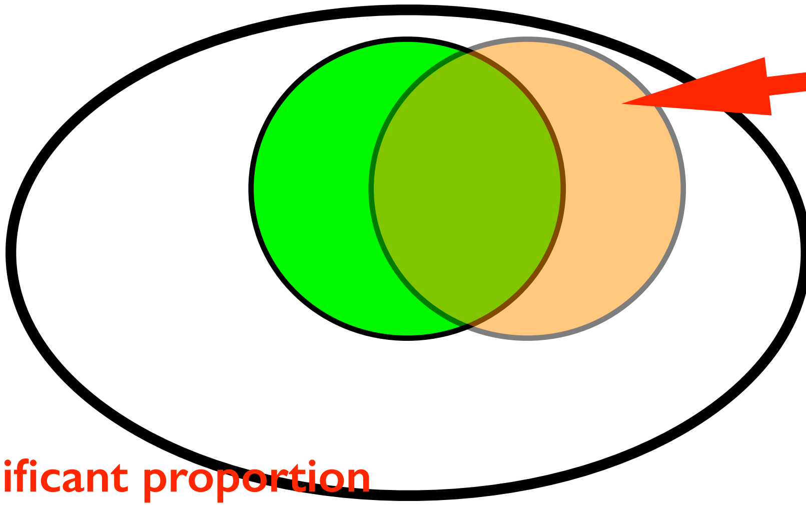
- unique contributions of variance above and beyond other variables
- problem: we can still end up with variables in the equation that don't account for a significant unique proportion of variance



significant proportion
of variance

add X1

$$Y = \beta_0 + \beta_1 X_1$$

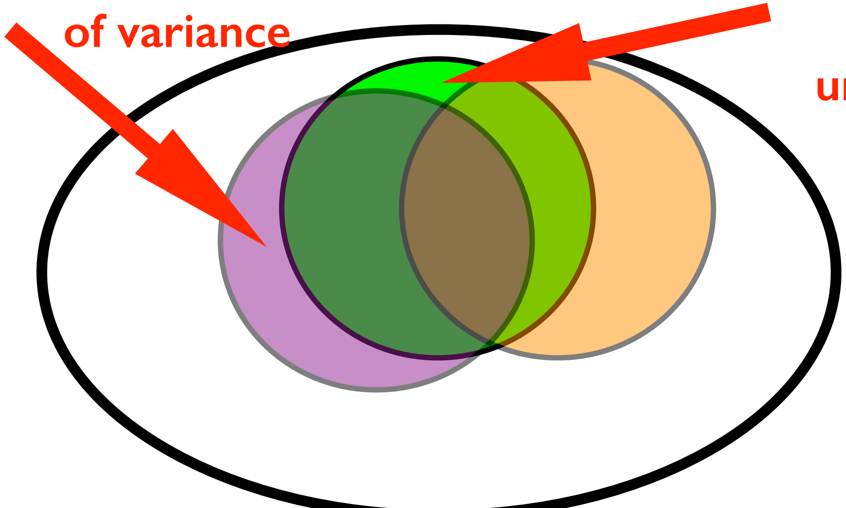


significant proportion
of variance

add X2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

significant proportion
of variance



PROBLEM!

no longer significant
unique portion of variance

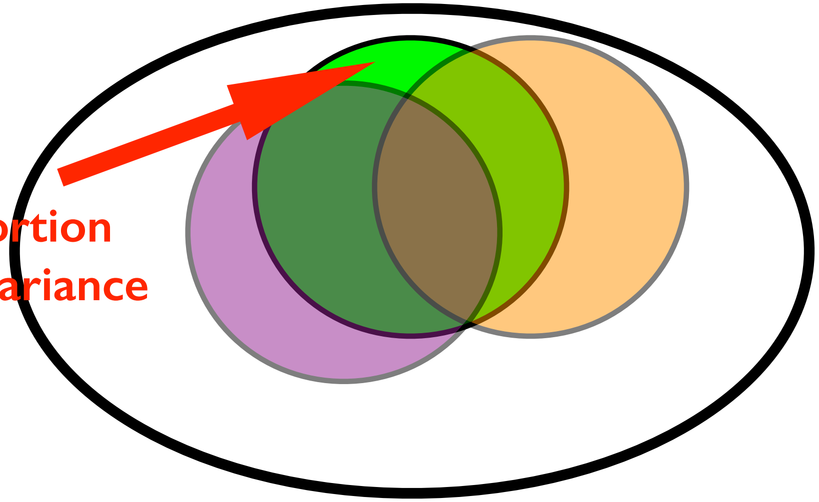
add X3

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Backward Regression

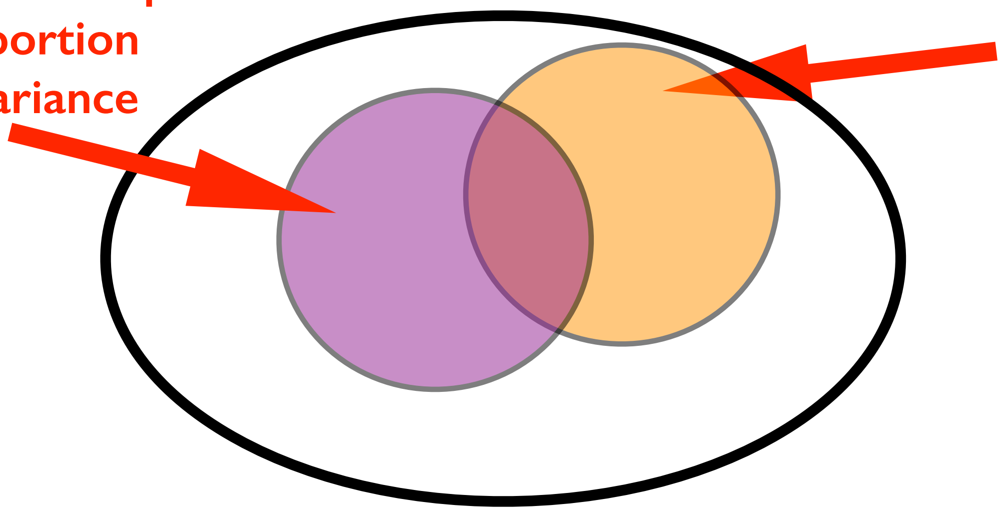
1. start with all IVs in the equation
2. check to see if any IVs are not significantly adding to the equation
3. if no, stop
   if yes, remove the worst IV (smallest r^2) and go back to step 2

- backward regression avoids the problem of ending up with variables in the equation that don't account for significant unique portions of variance



remove X1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

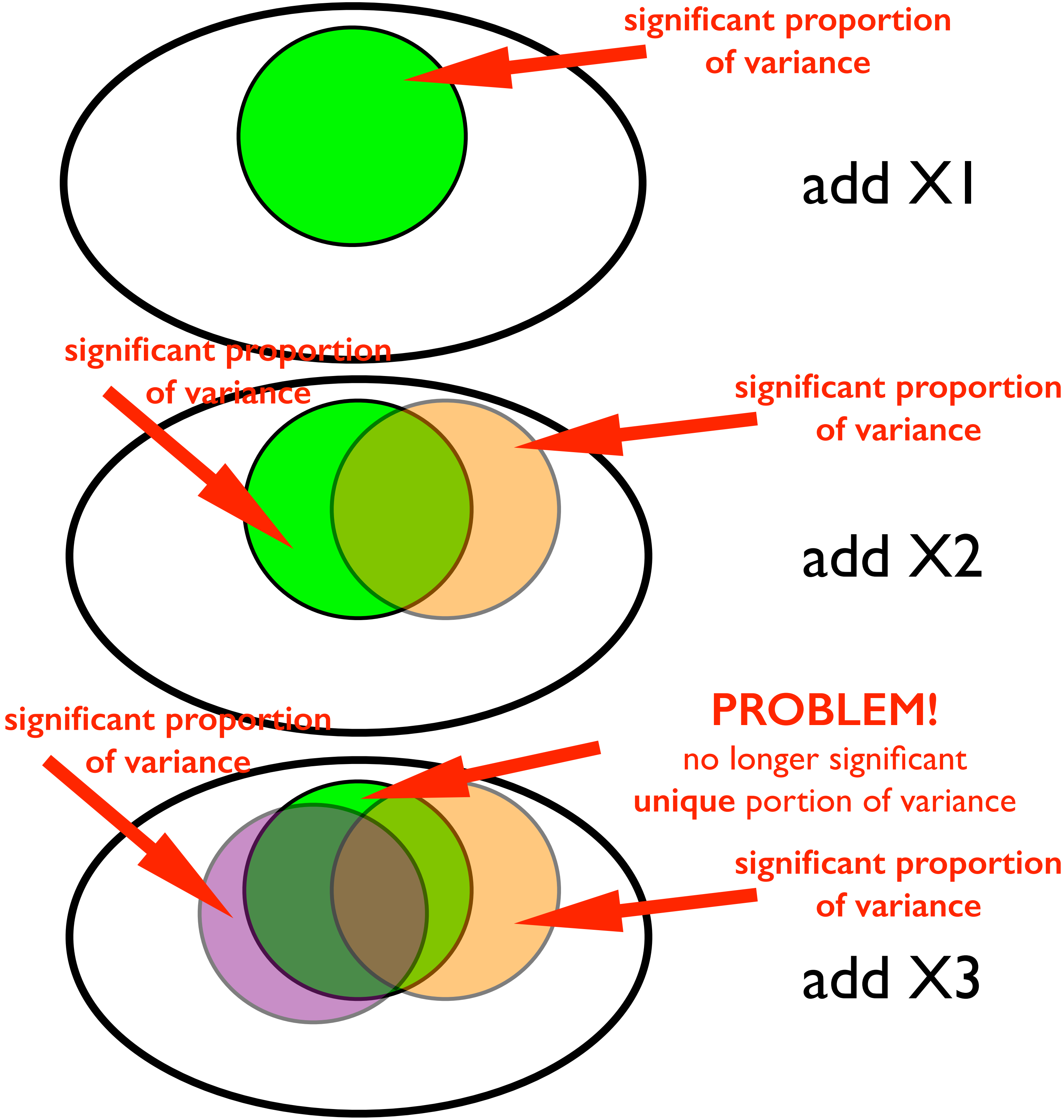


stop

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3$$

Stepwise Regression

1. start with no IVs in the equation
2. check to see if any IVs significantly predict the DV
3. if no, stop
   if yes, add best IV (largest r^2) and go to step 4
4. check to see if any IVs add significantly to the equation
5. if no, stop
   if yes, add best IV (largest r^2), go to step 6
6. check each IV currently in the equation to make sure they contribute unique portions of variance
7. **remove** any that don't
8. go to step 4



$$Y = \beta_0 + \beta_1 X_1$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

remove X1

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3$$

Building Models

- stepwise regression is almost exclusively used these days
- backward and forward regression not very common any more
- how to decide if a variable when added or removed is significant?
    - F-tests, using p-value cutoff (e.g. 5%) - this is how SPSS does it
    - Akaike Information Criterion (AIC) - another measure of the tradeoff between model simplicity and model goodness-of-fit (this is how R does it)
        - http://en.wikipedia.org/wiki/Akaike_Information_Criterion

Benchmarking

- when we ask the question "does variable X3 contribute unique variance" we are comparing one model against another
- this is known as benchmarking
- news-flash: we have been doing this all along!
- we are comparing a full model and a restricted model
- restricted: $Y = \beta_0 + \beta_1 X_1$
- full: $Y = \beta_0 + \beta_1 X_1 + \boxed{\beta_2 X_2}$
- F-test tests whether X2 adds **unique** variance over and above that already accounted for by the restricted model